Examensrapport

Wilhelm Kärde

November 16, 2014

Abstract

Contents

1 Introduction

Grammar in texts can nowadays easily be checked via word processors such as *Microsoft Word* [2], *Apache OpenOffice* [3] and even online spellcheckers such as *jspell* [4]. They are great tools to correct spelling and other common grammar mistakes. When writing texts however, whether it be a narrative, argumentative, expository or a descriptive text, there is more to what makes it a good text than its grammar. For instance they do not take into account what type of text one is actually writing. A well written argumentative text should reason for or against some matter and if it does not then it can not be considered a well written text despite having excellent grammar.

In schools, a grading matrix can be used by teachers in order to assess whether certain assignment goals have been reached when grading texts and it is useful for the student to understand what is required of their text in order to achieve a certain grade. The matrix typically consists of columns or rows of assignment goals, such as structure, content, language or similar on the one side and matching grades on the other. It is however up to the students and teachers themselves to assess whether or not the text they are writing or correcting reaches any of the assignment goals. This requires both groups to already have the linguistic skills in order to so, which can not be required by the student group.

The linguistic skills vary between students. While some may well be able to assess their own writing in order to obtain the highest grade, others will not. Students may even suffer from cognitive disabilities which further hinder their writing. This puts a lot of strain on teachers who have to deal with students needing variable degrees of help in their writing. Evaluating which students need the most help during the students' writing process is difficult, as it is hard for teachers to have a complete knowledge of how well all students are performing during any given moment.

The goal of this project is to develop a tool which helps students achieve their assignment goals through linguistic text evaluation. The tool should furthermore be of assistance to teachers assessing which students need the most attention at any given moment. This requires a real-time tool which feed-forwards information to both students and teachers. The texts are restricted to the Swedish language and students from grades 7 through 9.

2 Theory

2.1 Language technology

There exists today several different categories of language technologies associated with grammar checking, some of them will be introduced in this chapter. Focus lays on technologies which have been used for the Swedish language.

2.1.1 Rule based

A natural way of detecting grammar errors in a text is to translate the grammar rules in a language into rules that can be processed by a programming language. A tool that utilizes this technique for the Swedish language is Granska [9], developed at KTH (Royal Institute of Technology) in Stockholm. It uses a part-of-speech tagger to disambiguate all the words in the text. It then runs each sentence through a set of rules. The rules match the part-of-speech combinations of words in a sentence which correlates to some grammatical rule in the Swedish language.

In the Swedish language for example, for a sentence to be grammatically correct, there has to be a subject present. Thus one of the rules in the rule library states that for each word in this sentence, if not one of the corresponding part-of-speech tags is a subject, report that this sentence is incorrect.

A drawback with rule based NLP is that it requires extensive work. Each rule has to be hand written and take into account all the exceptions that may be associated with it. However, the feedback which it can generate is often very powerful as each rule accounts for a specific grammatical specification in the language. Furthermore, the recall and precision of the tool are easy to calibrate as rules can easily be added, removed or altered.

2.1.2 Statistical

A common strategy within statistical natural language processing (NLP) is to use n-grams. An n-gram is a table with frequencies tied to a number of *n-large* sequences of characters or words. Thus, a word 3-gram in English has entries such as:

freq	word1	word2	word3
40	like	a	chain
50	like	a	challenge
42	like	a	champ

Table 1: A word 3-gram for English

By processing large amounts of text from a specific language, the frequencies of common and uncommon sequences of words can be produced. Uncommon sequences are potential grammatical errors.

Instead of words, characters in words can be used. As with word n-grams, uncommon character n-grams are a potential spelling errors. However, they can also be the consequence of a rare occurrence of a word in the language and thus

a rare occurrence in the corpus.

ProbGranska[8] from the CrossCheck project uses PoS (part-of-speech) tag 3-gram frequencies. Furthermore, it attempts to address the problem where rare but grammatically correct word 3-grams are regarded as incorrect. It calculates a distance between tags and then attempts to replace the rare one with one that is more frequent and closer in distance, i.e has a more similar syntactic context. The switching of tags is penalized based on a probability of retaining grammaticality for the specific switch multiplied by the frequency of the newly formed 3-gram. If the penalized frequency is high, the newly formed 3-gram is probably correct and thus probably also the former 3-gram which was then most likely a rare but correct word 3-gram.

2.1.3 Machine learning

Various machine learning algorithms can be used to classify and thus identify different error types. They rely on big corpuses from which they learn how different classifications should be done and can then apply what it has learned by classifying new sets of texts.

An area in which machine learning is used extensively is within *Automated Essay Scoring* (AES). The algorithms used are trained with corpuses of student texts that are of varied quality. They then gather information about these texts based on a few criteria such as word variation and text length. What the algorithm then learns is that bad texts are associated with a certain word variation and a certain text length. It can then measure those values in new texts and classify them according to how they were classified in the corpus. This method of machine learning is called supervised learning, as you tell the algorithm what to look for and the texts are labeled.

An AES system for Swedish [14] has been developed at *Stockholms Universitet* in Stockholm. It used what they called 'simple features', 'language error features' and 'corpus-induced features'. The simple features were similar to those already discussed, such as text length and average word length. These could be directly measured from the texts. Uni- and bi-gram statistics were used to detect split compound errors. Along with a spell checker they made up the 'language error features'.

Lastly, statistics were gathered from sources such as blog posts and news articles. They were then compared to the student essays to obtain new variables that often correlated strongly with essay grades.

Performance evaluations of the system showed higher accuracy when assigning grades than the scoring accuracy between two humans. Additionally, the number of times the grades differed by more than one grade was lower than between two humans.

Many machine learning algorithms can also be used for more traditional grammar checking. By feeding the algorithms texts that are annotated with its grammar errors and its correct grammar it will become trained to detect errors in text. However, it requires a lot of work to annotate the texts, especially since machine learning algorithms usually require as many as possible.

A way around this, which was implemented in SnålGranska in the CrossCheck [8] project, is to automatically insert errors in (mostly) correct texts and then annotate these errors at the same time. However, these errors don't

necessarily reflect the way actual humans error when writing texts and so it is therefore limited in its quality as a grammar checker.

2.2 Writing aspects

Skolverket[1] has published *Språket på väg* [15] which is material aimed at helping teachers asses their students linguistic development and competence. It contains more fine grained information about which aspects are important for the students to learn in order to meet the requirements at the end of their ninth school year. Teachers are encouraged to use it as support in their evaluations and they are also encouraged to walk students through the aspect matrices which it contains.

These aspect matrices are divided into four categories: conversing, speaking, writing and reading. The only relevant matrix for this thesis, writing, contains 5 aspects (some of which are further broken down into sub aspects):

- *Recipient*
  -How the text works for intended readers
  -Communicative and cognitive functions
- *The text's genre*
- *Overall structure and language - global and local levels*
  -Textual level
  -Paragraph level
  -Sentence level
  -Expression and word level
- *Norms and language correctness*
  -Paragraph indications
  -Sentence constructions
  -Construction of expressions
  -Word usage
  -Spelling
  -Punctuation
- *The text in its entirety*

The matrix in its entirety can be found in appendix A.

[1] Swedens teaching institute for public education

2.3 Writing aspects from Skolverket and NLP

The complexity involved for a machine to measure the quality of the aspects introduced in section 2.2 vary greatly. The least complex aspects involve mostly syntax evaluations and the more complex ones involve more semantic evaluations of the texts. For some aspects, even the intentions of the writer is part of the evaluation. These aspects are more suited for humans to evaluate. To gain a deeper understanding of each aspect and how they relate to NLP, a brief walk-through of each will be the focus of this section.

2.3.1 Recipient

To properly evaluate this aspect, the semantics of the text have to be measured. For example, the sub-aspect: *How the text works for intended readers* involves evaluating how well the intentions of the text (to amuse, convince, inform, etc.) is communicated to the reader. While hard to extract the true semantic values from a text, there are methods which could be of use. Machine learning can be used to categorize the texts if a corpus is provided. It could thus be used to find patterns in the text which are specific for the different categories. Texts which are informing might use a different vocabulary than texts which are amusing for example.

Communicative and cognitive functions are measures of how well suited the linguistic language and stylistic choices are in relation to the purpose of the text. It also takes into account the logical construction of the text as well as its relation to potential sources. If provided, lexicons with appropriate and also inappropriate words could be used for different texts in order to evaluate the language and stylistic choices of a text. Statistics over frequent words, such as a word 3-gram, for texts with different purposes could be obtained and used as measurements as well. However, the precision of such approaches are likely to be low as there is a high probability for false alarms.

The logical construction of a text takes into consideration how clean-cut and coherent it is. The definition of what constitutes a clean-cut text is unclear and arguably subjective. But the texts could be measured on sentence and paragraph length for example to make sure they are not too long. Coherency is another aspect depending on semantics. Local discourse coherency, as defined by a measure of Centering Theory's [12] *Rough-Shift* transitions was evaluated in *Evaluation of Text Coherence for Electronic Essay Scoring Systems* [13]. The results of its use within AES were positive and proved to be a good measurement of short-lived and unconnected topics within paragraphs.

2.3.2 The text's genre

A text can be written in a variety of genres. Each genre follows a convention of writing. Skolverket has provided in *Språket på väg* [15] several criteria which should be met when writing *a letter*, *letter to the editor* and *a story*. To present the challenges involved with evaluating a genre, the criteria involved in writing a correct *a letter to the editor* will be presented. The full list for all three genres can be found in Appendix B.

The first criterion states that there has to be a thesis or a main thought and an argument in the text. A machine learning algorithm can be used to try and learn the structure of these kinds of sentences and thus possibly identify

them. However, the second criterion which looks at the quality of these is more difficult. The evaluation of for example the relevancy of an argument is not supported in any language technology to date. The same can be said about a number of the other criteria in this category: following the logic of the reasoning in the text, looking at its nuances, identifying commitment and emotions are all areas in which humans are more suited for evaluation than NLP.

Two criteria for which NLP is more suited are the ones involving tens and expressions expressing contradictions, emphasis and conclusions. Usage of a consistent tens can be verified using rule based technology and with the help of lexicons, desired expressions can be identified and matched in the text. However, evaluation of how well these are used in the text is another area where NLP is struggling.

2.3.3 Overall structure and language - global and local levels

The overall structure and language category stretches from a global textual level of evaluation to a local word and expression level. At the *textual level* texts are evaluated on disposition and on how well the title matches the rest of the text. While difficult to assess the quality of the disposition, a machine learning algorithm can be used to identify certain parts of the text. Evaluations can thus be made on which parts are present and in what order they are written. How well the title matches the rest of the text depends on semantics. However, keyword extraction can be of use to link the text to its title.

At a *paragraph level*, once again a semantic analysis on how well the paragraphs work, by themselves and as a whole, is hard to measure. However, *Rough-Shift* analysis within paragraphs, as introduced in 2.3.1, can be a good measurement on how coherent and tied together a paragraph is.

Working at a *sentence level* and focusing more on syntax suits NLP technologies well. To identify if a subject initiates each sentence or if main clauses in sentences are correctly separated with commas or conjunctions depends solely on syntax and the grammatical rules of the language. There are however evaluations on this level depending on semantics as well. The more advanced texts at this level have to incorporate a syntactical variation with a comfortable flow. Furthermore, main clauses that are not separated correctly can at this level be encouraged if they are stylistically on purpose. Measuring the comfort in a flow or analyzing the purpose of the writer is of course not a task in which NLP fares very well.

The last sub-aspect is at a *expression and word level*. Well written texts should vary the words used and the words and expressions used should work well with the rest of the text. Furthermore, the word choice includes whether the words used are too trivial and general as opposed to creative and thus enriching the texts. Different algorithms to measure the lexical diversity are technically possible, but the true semantic enrichment gained from diversity can presently only be measured by humans.

2.3.4 Norms and language correctness

As discussed in 2.3.3, whether a text consciously breaks stylistic norms or not is difficult to asses with NLP. However, this sub-aspect is primarily concerned with language correctness, in which NLP fairs much better. Especially since this

sub-aspect is primarily concerned with syntactical correctness. With *paragraph indications* writers should be able to be consistent in the way they separate paragraphs and they should do so with either a blank row in between them or with a tab to mark to beginning of a new paragraph.

With *sentence construction*, language correctness takes into account how complex the sentences are. A complex sentence in this context is composed of one or more sub-clauses and thus also often of more words than a non-complex sentence. Texts are encouraged to incorporate complex sentences as long as they are syntactically correct most of the time. The criteria are similar for the *construction of expressions*, only it is not as clear what a complex expression consists of, other than that it is long, complex and unusual. For this reason it can be hard to measure the complexity of expressions. Sentence complexity on the other hand is better defined and thus easier to measure and the paragraph aspect is simply a matter of parsing.

The *word usage* in this sub-aspect has more or less the same criteria as was used for the *expression and word level* from section 2.3.3. The difference being that expressions are not a factor and that the usage of words aspect deals with words that are considered hard and unusual as opposed to general and trivial from section 2.3.3. Statistics can be of use to both make sure that some words used are in general infrequently used but that they at the same time are used in a correct context via the use of n-grams.

Lastly there is the *spelling* and the *punctuation* aspects. Spelling is based on whether there are systematic errors and also if the spelling is correct with regards to the genre the text is aiming for. If specifications can be provided as to how the spelling differs for different genres, both of these criteria can be checked for with NLP technologies. As for the punctuation aspect, the criteria are similar to other aspects. the syntactic use can be checked but the semantic use of how well for example the commas in a text are used is hard to measure.

2.3.5 The text in its entirety

This aspect is intended to be a measurement of far the student has reached in its writing abilities. Thus it takes into account all other aspects and summarizes how well the student's performance is overall. It also looks at a historical perspective on the students writing in order to show whether the student is improving or not.

3 Method

For this thesis I have collaborated with Invigos AB. They have provided access to student texts, experts within school evaluation and a platform to develop, integrate and test the tool developed in this thesis. Together with Invigos AB and KTH we elected to use Granska, introduced in section 2.1.1, to scrutinize and give feedback on student texts. A rule based technology has the advantage of being able to give fine grained feedback and does not have any dependencies towards big corpuses of annotated student texts, which there is not an abundance of because of copyrights and due to the fact that there are too few Swedish speakers.

3.1 Custom aspect matrix

Since it is not possible to cover all aspects introduced in section 2.2 with a rule based solution, a custom aspect matrix was developed together with Steve Wretman and Helena Moreau, who both have been working together with Skolverket on the subject of evaluation. This matrix, seen in figure 1 below, represents which quality aspects of written text are relevant and also possibly measurable with rule based technology.

As can be seen in the matrix, next to the column of aspects are three more columns: *Påbörjad* (begun), *På God Väg* (well on the way) and *Behärskar* (masters). The idea is to give students a status on each of the aspects. A criterion to get the status on an aspect is that the text triggers rules in Granska associated with that particular aspect. More details about how this level is set is discussed in section 3.3.1.

3.2 Granska

Part of this project involved publishing Granska under the "GNU GENERAL PUBLIC LICENSE, Version 3, 29 June 2007". With that, it was necessary to have a single version working under both Solaris and Linux, 32- and 64-bit. The main challenge in that work consisted of updating Granska's usage of the library Xerces, which was only compliant with its first major release. Xerces has since then been updated to release 3.1.1, which Granska is now compliant with. Other updates included merging the 32-bit version of Granska with its 64-bit version, which was a simple matter of updating a couple of data structures. Lastly, the Makefile had to be updated to take into consideration which platform it was being built on as Solaris and Linux in some cases uses different system libraries and building tools. This version is now available at GitHub [5].

An addition to the repository was the implementation of a small API with only two calls, one load function and one scrutinizing function. This enables programs to call the Granska API once to load and initialize the necessary resources which Granska depends on and then be able to call a scrutinizing function repeatedly with new data and thus be better suited to handle multiple requests.

Of course the main task has been to upgrade Granska's rule library to match the different aspects from the custom aspect matrix. To ensure that the rules implemented follows the actual rules and conventions of the Swedish language

Språkriktighet och grammatik Aspekt/innehåll	Påbörjad	På God Väg	Behärskar
Stavning Sje-ljud Tj-ljud Dubbelteckning e/ä o/å J-ljud Ng-ljud Suffix Prefix			
Skiljetecken Punkt Komma Frågetecken Utropstecken Kolon			
Stor bokstav			
Särskrivning			
Tempus Presens Preteritum (perfekt och pluskvamperfekt)			
Pronomen Personliga (including de-dem) Posessiva			
Dialog Talstreck Citattecken Skilja skriftspråk-talspråk			
Styckeindelning			
Huvudsats Bisats *Sambandsord*			
Adverbial Rumsadverbial Tidsadverbial			
Verb- Regelbundna Oregelbundna Huvudverb hjälp			
Beskrivningar Person Miljö Känsla			
Variation Inledning av meningar Längd på mening Sambandsord Adjektiv Verb Ordval			
Syftningsfel			
Var – vart			
Korrekt ordföljd			
Prepositioner			

Figure 1: Custom aspect matrix.

I have tried to make them follow the grammar described in the following literature: *Svenska Skrivregler* [7], *Grammatik* [17], *Språkriktighetsboken* [6] and *Svenska Akademiens Grammatik, 2 ord* [16]. I have also had access to teacher evaluated student texts which have inspired new rules and served as a testing platform.

3.2.1 The Granska rule language

Granska has its own rule language built into it. It is too complex to be covered in this thesis and more importantly it has already been covered extensively[1] by Ola Knutsson. But for the sake of the readers of this thesis, an introduction of it will be presented here as it is important to understand both its strengths and weaknesses.

For each sentence in a text, a subset of the rules are triggered and applied to it. The role of the rule is to match a part-of-speech pattern within a sentence and then either accept it as correct grammar, mark the incorrect grammar and perform some editing to correct it or even re-tagg the whole sentence.

The syntax of a rule is as follows:

```
name@category
{
variablename1(),
...
variablenameN()
- ->
optionalmethod1()
...
optionalmethodM()
action()
}
```

The syntax can be broken up into three parts: header, text matching and action methods. The header consists of some general information about the rule: a name and a category. The name is simply a name unique to its category whereas each category has a structure associated with it as follows:

```
category name
{
info("Information about this rule, how it is triggered for instance")
link("URL to further reading regarding this rule")
}
```

The second part (above - ->) matches part-of-speech patterns in a sentence. Within the parenthesis of each method (which matches a word or a token), arguments can be passed to match not only part-of-speech specifics but also punctuation and even String matching. Furthermore logical operators such as & (and) and | (or) are available to account for multiple possibilities. A sequential variable can optionally be put after a clause to denote a repetition of the matching. The possibilities are: *zero or more*, *zero or one*, *one and more*

and lastly *one up to a maximum of N*.

Granska uses a stochastic tagger [10] which is based on a Markov model of the language. The tagger disambiguates the words and associates with each word a tag that describes its part-of-speech and morphological features. This information is available in the rule language through a characteristics class. Examples of attributes in this class is: wordcl (word class), gender (gender), sed (sentence delimiter) and vbf (verb form). These attributes are set to specific values by the tagger. A verb in the text may thus be set to wordcl=vb (verb) and vbf=prs (present tense).

Lastly in the third part (below - ->), it is possible to perform actions based on the matching found previously in the rule. The only mandatory action is the call to the *action* method, which states what kind of rule it is. A rule can be accepting, scrutinizing, search or editing, re-tagging or a help rule. Optional methods include, among others, *mark* (marking of desired objects), *jump* (jump to a location in the rule file) and *corr* (edit the objects in some way). In order to edit the variables there are a number of methods available, including *delete*, *replace* and *substr* to name a few.

An important feature of Granska is its integration with Stava [11] which offers a spelling error detection and correction API towards Granska. Stava was especially developed to handle compound words and inflections in Swedish by using a probabilistic method and a word list stored as a Bloom filter. Two important functions within the rule language is: *spell_OK* and *spell_corr* which checks if a token is correctly spelled and returns correction suggestions respectively.

An existing rule in the rule library which looks for a spelling mistake is shown below:

```
stav1@stavning
{
X(wordcl!=pm & !spell_OK(real_text, token))
–>
corr(X.spell_corr(X.real_text, X.token))
info(spell_info(X.token))
action(scrutinizing)
jump(slut_stavning, 0)
}
```

The word class *pm* stands for proper noun. So in short the rule above triggers on any word that has a word class tagg other than a proper noun and where *spell_OK* reports that the word is incorrectly spelled. If it triggers, spelling corrections of the word are returned by *spell_corr* and it then jumps to *slut_stavning* which is located after all the remaining rules belonging to the category *stavning*.

3.3 Granska-service

As previously mentioned, I have collaborated with Invigos AB in this project. They are developing a tool that is to be used in schools and in this tool they intend to use Granska. As their goals coincide with the goals of this thesis, some

of the work involved integrating Granska with their Granska service as well as developing on the service itself.

3.3.1 Mastery

A common goal was to give feedback to the students based on the aspect matrix from section 3.1 and that meant showing not only what is bad within a text but also what is good. As it was however, Granska only outputed errors and information about those errors to the end user. Therefore, we elected to change Granska so that it also outputs which rules were triggered in a text and how many times they were triggered.

It is necessary to say a few words on how a rule is triggered in Granska as not all the rules are triggered for each sentence. An optimization is made beforehand. For each rule, a least possible tag bi-gram is computed and stored. This least possible tag bi-gram is the bi-gram consisting of the two consecutive positions in the rule that have the rarest two combining tags when compared with bi-grams of the Swedish language. Furthermore, all possible words are also computed for each position in the rule. Two tables are thus created, one that maps tag bi-grams to rules and one that maps words to rules. When Granska is run on a sentence, it maps all its tag bi-grams and words to the rules in these two table and runs only these rules on that sentence.

As Granska now outputs how many times each rule triggers, it is now possible to combine the metrics of the triggered rules and the errors produced. This new metric we named *mastery* and is the basis for how the status of each aspect was computed.

3.3.2 Post-processing

Some of the aspects from section 3.1 does not suite Granska and a rule based approach quite that well even though we had Granska in mind when choosing them. We included them on the basis that it was possible to make at least a simple evaluation of them by post-processing the text.

The output from Granska is in the form of XML. It begins with the information about which rules that were triggered and then follows information about each sentence, including every words part-of-speech and morphological features. After each sentence there is additional information in the case that the sentence contained any errors. This structure allowed for the post-processing to be done at the same time as the processing of the rest of the output from Granska and is performed on the service.

4 Results

The results are presented in two parts. First, a walk-through of the maturity of the aspects from section 3.1 is presented and then follows a comparison between the output of Granska and teacher evaluated texts. An overview of the results of this part is depicted in figure 2 and 3 on the following pages.

4.1 Maturity of aspects

4.1.1 Feedback on sound-post errors

Granska has been extended to give feedback when an error is related to one of the different sounds: sje, tje, e/ä, o/å, j and ng. All these sounds belong to its own category where it is pronounced in one way but it is spelled differently in different words. For example, in the category *j* there are 6 different combination of letters, each making up the sound *j* in different ways: j, g, lj, dj, hj, gj. The words *jordgubbar* (strawberries) and *hjortron* (cloudberries) are both pronounced with an opening j-sound only there is a silent h in *hjortron*.

If *hjortron* were to be spelled as *jortron* instead, Stava would recognize that *jortron* is not a word and it would also recognize that the first letter *j* belongs to the sound class of *j*. It would then substitute the *j* with each of the different letter combinations belonging to the sound class *j*. This would in turn create the new words: gortron, ljortron, djortron, hjortron and gjortron which are further spelled checked, resulting in only *hjortron* passing the test.

A new method *spell_OK_sound* has been added to the Granska rule language. It is a boolean function which via Granska makes a call to Stava to check if a word is misspelled with regard to any sound error, as previously explained. At this stage the substitutions are performed but not stored as any suggestions. The reason is that it is possible that there could be more suggestions available than the sound substitutions. The misspelled word *helj* for example, is substituted with *helg* by the sound-post algorithm but other non sound related suggestions are also presented by Stava, including: hela,helt,hel, hej and hell. Therefore the rule utilizing the *spell_OK_sound* function makes a call to *spell_corr* which returns all possible suggestions. Thus an error which is classified as a sound error is in fact an error in which one sound replacement is possible. Note however that this does not trigger if the error is unrelated to a sound post. The misspelled word *självkänlsa* does not trigger the sound post rule as the ending *lsa*, which is corrected to *lsa*, is not related to any sound post replacement.

Stava had support for the categories sje and tje beforehand while the other categories have been added. The sound substitution is part of Stava's correction algorithm and still is. The new categories are of course an addition and a new API call to Stava has been added which runs the same sound substitution algorithm but only returns a boolean instead. This way Granska can call the boolean function to check for sound substitution to be able to categories the error and it can then perform a regular spelling correction call to Stava.

4.1.2 Spelling, split compounds and agreement errors

A number of aspects could easily be mapped to categories in the rule library of Granska without any further extensions to the library. These included *Stavning* (spelling), *Särskrivning* (split compounds) and *Syftningsfel* (agreement errors).

Maturity	Aspect
"Dubbelteckning" is not supported as it causes too many spelling errors to associate with this category.	Feedback on spelling errors with respect to sound-posts: Sje-ljud Tj-ljud Dubbelteckning e/ä o/å J-ljud Ng-ljud
Spelling is corrected well by Stava.	*Stavning*
Granska has been extended to support skiljetecken.	*Skiljetecken* • Frågetecken • Utropstecken
Capitalization has been improved in Granska.	Stor bokstav • Egennamn • Ny mening • Felaktig versal
Compound words are well supported within Granska with the help of Stava.	Särskrivning
A complex aspect that has been improved in Granska, mainly with respect to the sub-aspect: De-dem	*Pronomen* • de-dem • Objekt/subjekt
Not supported.	*Dialog* • Talstreck • Citattecken • Skilja skriftspråk-talspråk
Good paragraphing depends on semantics which is out of scope for Granska. Coherency in the syntactical sense depends on information of it beeing available.	Styckeindelning
Complex aspect which proved to be out of scope for this project. Further resaearch is encouraged.	Huvudsats Bisats *Sambandsord*
The semantical usage of this aspect is out of scope for Granska.	*Adverbial* • Rumsadverbial • Tidsadverbial
New category that has been introduced in Granska. Errors in this aspect are rare among native speakers.	Komparation • Suffixkomparation • Perifrastisk komparation

Figure 2: Overview of the resulting aspect matrix, part 1. Green represents the supported aspects and red the unsupported ones.

Granska supports the correction of grammatical mistakes involving verbs. It has been extended to support correction in the misuse of tempus-shifting within sentences. The granularity to give feedback on the type of verb is not supported however.	*Verb* • general verb issues • Regelbundna • Oregelbundna • Huvudverb • hjälp • tempusskifte
Can't diffentiate which adjectives are descriptive specifically for Person , Miljö or Känsla. Naive implementations on Variation: Word choice too semantic.	*Beskrivningar* - Person , Miljö, Känsla *Variation* • Inledning av meningar • Längd på mening • Sambandsord • Adjektiv • Verb • Ordval
Well covered within Granska.	Syftningsfel • Predikativ • Övriga
Need a semantical knowledge of direction and place which is out of scope for Granska.	Var – vart
There is support in Granska. However, this aspect is very complex and hard to cover with a rule based technology. It is also unusual to come across errors from native speakers within this aspect.	Korrekt ordföljd • upprepning • ordföljd i bisats • ordföljd i indirekt frågesats
The semantical usage of this aspect is out of scope for Granska.	Prepositioner

Figure 3: Overview of the resulting aspect matrix, part 2. Green represents the supported aspects and red the unsupported ones.

These represent the most important categories in Granska and as a lot of effort already has been put into developing them, no work has been done in this project to extend these any further.

4.1.3 Skiljetecken

Several categories were either created or extended. One of the new categories is *skiljetecken* (punctuation). It has two sub-categories associated with it: *Frågetecken* (question mark) and *Utropstecken* (exclamation mark). Originally however, this aspect had 5 sub-aspects, including: *punkt* (period), *komma* (comma) and *kolon* (colon), as can be seen in figure 1 from section 3.1.

They are omitted for two reasons, and the first reason is that there are little to no rules in the Swedish language regulating the use of these. The other reason being that they are to an extent incorporated into other rules or aspects. For instance, both *period* and *comma* are closely related to the aspect *Variation-längd på mening* (sentence length). In the case where a sentence is very long, it often contains several commas of which one or several most likely should be a period instead. The use of the commas does not brake any grammatical rules, but in general and especially in school texts the advice is to keep sentences relatively short to allow for a better reading flow for the reader. It is impossible for a rule based technology to give any assessment on reading flow and so that task is left to the writer with the help of the detection of long sentences by *Variation-längd på mening*.

Other categories to which *punkt* is related to are *Frågetecken* and *Utropstecken*. The reason being that they are both triggered when a period ends a sentence that most likely is either a question or an interjection. The way in which these rules work is that they look for words at specific positions in a sentence. For example, a sentence starting with a so called *frågeord* (question word) is probably a question. It is difficult to start a sentence with the word *why* and steer it towards not being a question. Similarly, if a sentence starts with a verb in the imperative tense it is probably an exclamation and should therefore end with an exclamation mark. The rule looking for a question word is shown below:

```
frågeordsfråga@frågetecken
{
X1(sed=sen | cht=mid),
X2(text="vem" | text="vad" | text="vart" | text="var" | text="varifrån" |
text="vilka" | text="vilken" | text="hur" | text="varför" ),
X(wordcl!=pn & text!="som"),
X3()*,
X4(text="." | text="!"),
X5(sed=sen)
–>
mark( X1 X2 X X4)
corr(X4.replace("?"))
info("Fråga?")
action(scrutinizing)
}
```

There are two things worth noting in the rule above. First of all, the rule actually checks whether the *frågeord* is at the beginning of a sentence (sed=sen) or if it follows a comma (cht=mid). This means that a misplaced comma triggers this rule which illustrates once again just how close the aspects within punctuation are.

The second thing to note is that the rule looks for words following the *frågeord* which nullifies the *frågeord's* effect of turning the sentence into a question. Both pronouns and the relative pronoun *som* have the effect of turning the sentence into either an indirect question or a statement instead.

4.1.4 Comparative adjectives

Comparative adjectives correspond to the aspect *Komparation*. Adjectives that are of a comparative nature have three inflection forms: *positiv*, *komparativ* and *superlativ*. In this aspect the focus is on errors inflecting in the *komparativ* form. They can be inflected in one of two ways, either with a suffix or with the addition of the modifier *mer(a)* or *mest* in front of it. In the case where it is inflected with a suffix it is a *suffixkomparation* and otherwise it is a *perifrastisk komparation*. Adjectives that are inflected with a *perifrastisk komparation* are often categorized as *long* adjectives as they are often words of a more complex nature with characterizing endings such as: *-(a)nde*, *-ende*, *-ad* and in most cases also *-(isk)*.

The rules that make up this category identify when the wrong inflection form is used for an adjective, that is when a long adjective is inflected with a suffix instead of a modifier and vice versa for adjectives that are not considered long. As an example, the word *lik* (similar or alike), is an adjective wich is inflected with a *perifrastisk komparation*. One would say: *Han är mer lik honom* and not *Han är likare honom* which would be the *suffixkomparation* equivalent. Other rules also try to indentify when a *komparativ* form of the word is missing altogether by searching for the word *än* (than) following the adjective. These types of errors are however rare among native speakers as they usually sound awkward when misused.

4.1.5 Pronouns

Pronouns is a large and complex aspect. The complexity is high because pronouns can refer to subjects or objects across sentence boundaries. Furthermore, it is often hard simply to disambiguate whether the pronoun itself is a subject or an object in the context.

In Granska, pronoun resolution across sentence boundaries is not something that is possible as Granska scrutinizes one sentence at a time. However, it would be possible to implement and analyse *Rough-shift* transitions in a text at the post-processing stage. No such implementation has been implemented in this project though, one major obstacle being that there are no paragraph delimiters in *txt* files. Because of this, the tagger of Granska can not disambiguate between paragraphs and thus there is no output information for the service to disambiguate between paragraphs either.

The problem of disambiguating a pronoun from a subject or an object is also a complex problem. In the Swedish language this is especially true for the

pronouns *de* (they) and *dem* (them). An underlying problem is that in the spoken language they are pronounced the same, unlike their English counterparts. For this reason, grammatical errors concerning *de* and *dem* are quite common in Swedish texts and it is why they have their own sub-aspect among pronouns.

Preferably, the tagger in Granska would do the disambiguation. Unfortunately it does its tagging based on the syntax, meaning that *de* is always tagged as a subject and *dem* as an object.

There existed some rules for *de* and *dem* in Granska before this project but it has been extended with a few more generating a better recall than before, whilst sacrificing some precision. The lowered precision is due to the added rules being based on assumptions and will in some cases fail. For example, one added rule looks at the position following the word *dem* and reports an error if that position is occupied by a verb. There is no grammatical rule that enforces *dem* to be written as *de* if it is followed by a verb, but it is a more common sentence construction in texts and examples of sentence constructions with a subject followed by a verb come easier to mind than one where an object is followed by a verb. All added rules are intolerant towards the spelling of de/dem as *dom* as it was not tolerated by the teacher in the teacher evaluated texts.

4.1.6 Capitalization

The capitalization aspect has been extended to support a few new sub-categories, namely: *Egennamn* (proper noun), *ny mening* (new sentence) and *felaktig versal* (bad capitalization). It is a relatively small aspect which relies heavily on the tagger being able to correctly tag each proper noun in the text. As the Granska API offers the method *is_cap* as well as the identification of a proper noun via the word class *pm*, the rules become very straightforward. The rule below identifies each proper noun that is not capitalized, it then replaces the word with its capitalized equivalent:

```
felgemen@egennamn
{
Y(!is_cap & wordcl=pm)
–>
mark(Y)
corr(Y.replace(firsttoupper(Y.real_text)))
info("Egennamn")
action(scrutinizing)
}
```

4.1.7 Verb

The verb aspect has been extended to include a *tempus* sub-aspect. It detects the switching between the past and present tense in a sentence, something that in most cases should be avoided. Thus, the rule searches for switches between the *preterium* (past) and *presens* (present) tense. The Swedish language does not have an explicit tense to mark future events which is why the rule is restricted to only the past and the present.

The *tempus* sub-aspect is not as applicable on texts outside the school spectrum which this project has focused on. But for student texts, where the writer in many cases unconsciously changes tense mid-sentence, it can give valuable feedback. It should be clear however that changing tense mid-sentence does not break any grammatical rule and is in fact sometimes necessary for the text to make sense semantically.

4.1.8 Variation

As Granska scrutinizes one sentence at a time and the variation aspect in some cases needs to take the whole text into consideration, this aspect has been implemented as post-processing on the service. The implementations are naive but easily adjustable by manipulating a variable. The list below shows the sub-aspects and how they are triggered, in parenthesis is the number currently implemented and used when testing:

- Inledning på meningar (Beginning of sentences):
  Triggers when X (3) consecutive sentences start with the same word.
- Längd på mening (Sentence length):
  Triggers when a sentence is longer than X (40) words.
- Sambandsord (conjunctions):
  Triggers when the same conjunction occurs more than X (2) times in a sentence.
- Adjektiv (Adjective):
  Triggers when an adjective is representing more than X (15) percent of the total amount of adjectives in the text.
- Verb (Verb):
  Triggers when a verb is representing more than X (15) percent of the total amount of verbs in the text.

4.1.9 Too semantic

A number of aspects depend too much on semantics in order to be evaluated with Granska. The use of *adverbial* (adverbial) and *prepositioner* (prepositions) are challenging in the same sense when it comes to semantics. The correct use of these depend on the context in which they are used. For example, *Han hängde upp tavlan i väggen* (He hung the painting in the wall) is most likely wrong and the preposition *i* (in) should be replaced with *på* (on). There is nothing grammatically wrong in that sentence, in some sentences the adverbial *i väggen* (in the wall) even makes more sense. NLP and particularly rule based NLP can at the moment not make that distinction.

Another distinction which is hard for NLP to solve is the one between *var* (where, referring to a location) and *vart* (where, referring to a direction). With a rule based tool such as Granska, where the only information available are the POS-tags of each word, there is not enough information to establish if a sentence is referring to a location or a direction. The two sentences: *Vart åker tåget?* (Where does the train go) and *Var står tåget?* (Where is the train) have the same POS-tag composition but the first is referring to a direction whilst the second one is referring to a location.

4.1.10 Too complex

An effort was put in towards covering the identification of *huvudsatser*, *bisatser* and the use of *sambandsord* (conjunctions) between them. A goal was to be able to detect *Huvudsatser* connected without the use of conjunctions, a so called *satsradning* in Swedish. The identification of *bisatser* would be helpful in evaluating the sentence complexity of a text. Unfortunately though, the structure of *huvudsatser* and *bisatser* is very complex, at least in the variation shown in student texts.

Rules were implemented and tested to identify *bisatser* following the scheme found in [6]:

bisatsinledare	subjekt	satsadverbial	finit verb	verbpartikel	objekt	adverbial

Table 2: Structure of a bisats from [6]

The performance of these rules on student texts were however too poor to be able to give any relevant feedback and were thus omitted from the final version of the Granska rule library. They are however still available, commented out, and further research on this topic is encouraged. As the use of clauses is a complex matter, a linguistic expert is encouraged to aid in the development of these rules. It should also be noted that it is possible that it is easier to write clause rules towards texts that are of a better linguistic quality than that of student texts.

4.1.11 Out of scope

A couple of aspects and sub-aspects were down prioritized and in the end fell out of the scope of this thesis project. The *Dialog* (dialog) and *Styckeindelning* (paragraphing) aspects were down prioritized as they are more a matter of parsing and of the identification of special tokens than one of natural language processing. Furthermore, the traits within these aspects that has to do with NLP, *Skilja skriftspråk-talkspråk* (separate speaking language from written) and *Good paragraphing*, are better solved using a machine learning approach.

Lastly, a couple of sub-aspects also fell out of scope. The disambiguation of associating a *general verb issue* with what verb type it concerns (*regular*, *irregular*, *main* or *help*) would need foremost the support from the Granska tagger to be able to disambiguate between them (currently only help verbs are). Even then it would be too time consuming to go through every rule related to *general verb issues* and analyse whether it can be disambiguated to concern only one type of verb.

In the post-processing aspect of *Variation*, support is not available for neither *Beskrivningar-Person, Miljö, Känsla* (Description-Person, Environment, Feeling) or *Ordval* (Word Choice). The former partly because it is already to some extent covered in the sub-aspect *adjektive* (adjective) of *Variation* and partly because there is no information from Granska that disambiguates between the three descriptions. *Ordval* (Word Choice) is simply too semantic to be covered with simple post-processing.

4.2 Mastery

As mentioned in section 3.3.1, information about how well a text performs on each aspect, called mastery, was implemented. Mastery has three levels: *Påbörjad* (begun), *På god väg* (well on the way) and *Behärskar* (masters). It is based on how many times a rule is triggered and how many times that rule reports an error. The information of how many times a rule is triggered was added to the ouput of Granska and the service calculates the mastery for each aspect.

Only a naïve approach calculating the mastery for each aspect was implemented. A more sophisticated approach where experts on evaluation would tune the mastery levels fell out of scope. The current method should be seen as a proof of concept.

In the current model, a level of mastery is only given on an aspect if it triggered at least 10 related rules. Note that triggering a rule only means that a rule has been called to scrutinize a sentence, not that an actual error has been reported. If an aspect does not trigger at least 10 rules its mastery is classified as UNKNOWN and no information about that aspect is shown to the user.

If an aspect has triggered enough rules, two treshold values are used: expert- and advanced-treshold. These tresholds are compared with the ratio between how many errors an aspect has triggered by how many sentences a text has. Thus, for an aspect to be considered to be at the level of *Behärskar* (masters), the service checks if:

$$nrOfErrors/nrOfSentences < expertTreshold$$

For *På god väg* (well on the way) the equivalent check is:

$$nrOfErrors/nrOfSentences < advancedTreshold$$

and if neither holds, it is considered to be at the *Påbörjad* (begun) level.

A test with the treshold values of 5% for the expert treshold and 20% for the advanced treshold was performed on 25 student texts (approximately 17000 words in total). It shows the spread of the mastery levels across the aspects. In most cases enough rules triggered to apply a mastery level. *Versal(Egennamn)* and *Skiljetecken(utropstecken)* were the only ones with a high UNKNOWN mastery, with 18 and 25 ocurrences respectively. The mastery level *Påbörjad* (begun) was most common in the *Pronomen(De/Dem)* aspect, with 10 occurences among the 25 texts. The complete spread is shown in table 3.

Aspect	Påbörjad	På God Väg	Behärskar	UNKNOWN
Pronomen(De/Dem)	10	3	11	1
Variation(Meningslängd)	7	10	8	0
Stavning(Stavfel)	5	10	10	0
Verb(Tempusbyte)	4	15	6	0
Verb(Verbfel)	3	9	13	0
Versal(Egennamn)	1	4	2	18
Kongruens(Kongruensfel)	1	9	15	0
Särskrivning	1	6	18	0
Variation(Sambandsord)	1	6	18	0
Versal(Gemen)	1	3	20	0
Versal(Ny mening)	1	3	21	0
Skiljetecken(Frågetecken)	0	3	22	0
Variation(Verb)	0	3	22	0
Kongruens(Predikat)	0	4	21	0
Variation(Meningsbörjan)	0	1	24	0
Ordföljd(I bisats)	0	2	23	0
Skiljetecken(Utropstecken)	0	0	0	25
Stavning(Ljudpost)	0	1	24	0
Pronomen(Object/Subject)	0	0	25	0
Ordföljd(Upprepade ord)	0	0	25	0
Komparation(Perifrastisk)	0	1	24	0
Variation(Adjektiv)	0	1	24	0
Komparation(Suffixkomparation)	0	0	25	0
Ordföljd(I indirekt fråga)	0	0	17	8
Granska/Service total	35	94	419	52

Table 3: Mastery spread across aspects, from 25 texts with approximately 17000 words in total. Expert threshold at 5% and advanced treshold at 20%.

4.3 Performance

Two tests were performed in order to evaluate the performance of the service using Granska. In the first test, 25 teacher evaluated texts (approximately 17000 words) from students of the same class (9th grade) were evaluated by the service. The output of the service was compared with the evaluations done by the teacher. Furthermore, 5 random texts were selected and were manually checked to locate any false alarm.

In the second test, a corpus of 10 plain texts (approximately 8000 words) from second language learners in the seventh grade were evaluated by the service and compared with the first test. Of those, 4 random texts were selected and were manually checked to locate any false alarm.

4.3.1 Comparison with teacher evaluated texts

The teacher evaluated texts are corrected with bold and blue highlighting and comments. The blue highlighting are words added to the text by the teacher

while the bold highlighting represents, in the teachers own words: *"Things that I have reacted to in the text. It can be spelling errors just as well as a compounding mistake, it can be the word choice, it can be text binding or something else."*

The comments vary a lot and are further reactions that the teacher has on the text. It is often in the form of a question or a suggestion. The questions generally question word choices and story related issues. The suggestions often have more to do with grammar and structure, such as paragraphing, punctuation and sentence length.

Since the teacher's corrections are not disambiguated by aspects, as Granska and the service is, no aspect-to-aspect evaluation is available.

An average of 22.4 errors per text were detected by the service to be compared with the combined average of 38.8 corrections and comments made by the teacher per text. This amounts to a total of 560 errors found by the service and 970 by the teacher. Most errors found by the service belonged to the *Pronomen(De/Dem)* aspect, mainly because the texts frequently make use of *dom*, which is not accepted by neither the service nor the teacher (in most cases). The two least verbose aspects, *Komparation(Suffixkomparation)* and *Ordföljd(I indirekt fråga)*, did not record any errors at all, The complete result can be found in table 4.

4.3.2 Recall and precision

It is difficult to assess the recall and precision of the Granska-Service solution. In order to calculate recall, all the errors have to be known beforehand. Furthermore, to evaluate each aspect, all the known errors also have to be classified by aspect. The teacher's corrections can unfortunately not represent all known errors as the teacher consistently misses errors and does not categorise its corrections the same way that the aspects are categorised.

The precision can be analysed by going through each error by hand, which of course is a time consuming task. Instead, five randomly selected texts have been analysed by hand in order to give a general idea of the precision to be expected for each aspect. In table 4, the result of this analysis is shown. In some cases, an alarm is triggered on a grammatical error caused by another kind of error. For example, the word "till exempel" is incorrectly abbreviated to "tex." instead of "t. ex.", causing a *Versal(nymening)* (new sentence) error. Even though the problem may be more related to *Stavning* (spelling) it is not considered to be a false alarm as it is still a grammatical error which needs attention.

The *Variation* aspect and all its sub-aspects can not generate false alarms as it simply counts different occurences in the texts and generates an alarm if the count reaches some treshold. All other aspects are generated by rules in the Granska rule library and are thus depending on the accuracy of the tagger and the rules.

The *Verb(Tempusbyte)* aspect had the highest number of false alarms, with a precision of 68%. The lowest precision was recorded by the *Pronomen(Object/Subject)* aspect with a 0% accuracy, with only a single error to compare with however. The second lowest precision (57%) was recorded by *Stavning(Stavfel)*, with 3 errors. The 3 errors are a consequence of the two words: *blogg* and *pusha* not being recognized. The total precision, with 110 errors found and 17 false alarms, was 85%. The complete result is available in table 5.

Aspect	Errors found	Average per text
Pronomen(De/Dem)	102	4.08
Variation(Meningslängd)	76	3.04
Verb(Tempusbyte)	68	2.72
Stavning(Stavfel)	56	2.24
Verb(Verbfel)	53	2.12
Versal(Egennamn)	33	1.32
Kongruens(Kongruensfel)	32	1.28
Särskrivning	24	0.96
Variation(Sambandsord)	21	0.84
Skiljetecken(Frågetecken)	12	0.48
Variation(Verb)	11	0.44
Versal(Gemen)	11	0.44
Kongruens(Predikat)	10	0.40
Versal(Ny mening)	9	0.36
Variation(Meningsbörjan)	7	0.28
Ordföljd(I bisats)	6	0.24
Skiljetecken(Utropstecken)	6	0.24
Stavning(Ljudpost)	3	0.12
Pronomen(Object/Subject)	3	0.12
Ordföljd(Upprepade ord)	2	0.08
Komparation(Perifrastisk)	2	0.08
Variation(Adjektiv)	1	0.04
Komparation(Suffixkomparation)	0	0.0
Ordföljd(I indirekt fråga)	0	0.0
Granska/Service total	560	22.4
Bold highlightning	273	10.92
Blue highlighting	43	1.72
Comments	327	13.08
Teacher's correction total	970	38.8

Table 4: Combined output from 25 texts.

4.3.3 Performance on second language student texts

No teacher evaluation is available for the second language student texts. Instead, the result is compared to that of the previous test. However, it should be noted that these texts are not only from second language students but also from students in grade 7, as opposed to grade 9 which was the case in the previous texts.

The result show an increase in the total amount of errors caught by the service. With an average of 43.6 per text it is almost twice that of the previous test (22.4). The most verbose aspect (92 errors found) is the *Versal(Gemen)* (lower case) aspect. This aspect reports an error when a word is wrongfully written with a capital letter. Three aspects all averaged around 6 errors per text, including *Stavning(Stavfel)*, *Pronomen(De/Dem)* and *Versal(Ny mening)*. A total of six aspects did not record any errors at all. Like in the previous test,

Aspect	Errors found	False alarms	Precision
Verb(Tempusbyte)	19	7	68%
Variation(Meningslängd)	18	0	100%
Pronomen(De/Dem)	13	1	92%
Verb(Verbfel)	10	2	80%
Särskrivning	7	2	71%
Stavning(Stavfel)	7	3	57%
Versal(Egennamn)	5	0	100%
Kongruens(Kongruensfel)	6	0	100%
Variation(Sambandsord)	6	0	100%
Versal(Ny mening)	4	0	100%
Skiljetecken(Frågetecken)	4	1	75%
Variation(Verb)	3	0	100%
Skiljetecken(Utropstecken)	2	0	100%
Kongruens(Predikat)	1	0	100%
Variation(Meningsbörjan)	1	0	100%
Stavning(Ljudpost)	1	0	100%
Pronomen(Object/Subject)	1	1	0%
Ordföljd(I bisats)	1	0	100%
Versal(Gemen)	0	0	N/A
Ordföljd(Upprepade ord)	0	0	N/A
Komparation(Perifrastisk)	0	0	N/A
Variation(Adjektiv)	0	0	N/A
Komparation(Suffixkomparation)	0	0	N/A
Ordföljd(I indirekt fråga)	0	0	N/A
Granska/Service total	110	17	85%

Table 5: Error analysis from 5 random texts.

neither *Komparation(Suffixkomparation)* nor *Ordföljd(I indirekt fråga)* recorded any errors, and neither did *Komparation(Perifrastisk)*, *Ordföljd(Upprepade ord)*, *Ordföljd(I bisats)* nor *Skiljetecken(Utropstecken)*. The complete result is shown in table 6.

4.3.4 Precision on second language student texts

Standing out in the result is the 0% precision of the aspect *Versal(Gemen)*, which also is the most verbose with 41 errors found. The low precision is however easily explained. The texts were themed around politics, and the names of the political parties were spelled with a capital letter in the texts. Granska however did not recognize these as proper nouns and thus reported an error on each occurence. All 41 of the errors were triggered by the capitalization of the name of a political party (around eight names).

The *Verb(Tempusbyte)* had an even lower precision (22%) than in the previous test (68%) and *Skiljetecken(Frågetecken)* recorded a 0% precision, albeit out of only 2 possible errors.

Among the better performing aspects were *Stavning(Stavfel)* with a precision of 94% (out of 31 errors) and *Pronomen(De/Dem)* with a precision of 81%

Aspect	Errors found	Average per text
Versal(Gemen)	92	9.2
Stavning(Stavfel)	63	6.3
Pronomen(De/Dem)	60	6.0
Versal(Ny mening)	59	5.9
Verb(Verbfel)	35	3.5
Verb(Tempusbyte)	32	3.2
Variation(Meningslängd)	16	1.6
Kongruens(Kongruensfel)	16	1.6
Särskrivning	13	1.3
Skiljetecken(Frågetecken)	13	1.3
Variation(Sambandsord)	9	0.9
Stavning(Ljudpost)	7	0.7
Versal(Egennamn)	6	0.6
Pronomen(Object/Subject)	4	0.4
Variation(Verb)	3	0.3
Kongruens(Predikat)	3	0.3
Variation(Meningsbörjan)	3	0.3
Variation(Adjektiv)	2	0.2
Skiljetecken(Utropstecken)	0	0.0
Ordföljd(I bisats)	0	0.0
Ordföljd(Upprepade ord)	0	0.0
Komparation(Perifrastisk)	0	0.0
Komparation(Suffixkomparation)	0	0.0
Ordföljd(I indirekt fråga)	0	0.0
Granska/Service total	436	43.6

Table 6: Combined output from 10 second language student texts.

(out of 26 errors). The texts frequently missed to capitalize the first word of every sentence, resulting in 27 errors and a 100% precision rate for the aspect *Versal(Ny mening)*.

The overall precision lands at 67%, with the aspect *Versal(Gemen)* bringing the percentage down significantly, although *Versal(Ny mening)* in turn helps raise the number. The complete result is shown in table 7.

Aspect	Errors found	False alarms	Precision
Versal(Gemen)	41	41	0%
Stavning(Stavfel)	31	2	94%
Versal(Ny mening)	27	0	100%
Pronomen(De/Dem)	26	5	81%
Verb(Tempusbyte)	9	7	22%
Variation(Meningslängd)	6	0	100%
Verb(Verbfel)	6	0	100%
Stavning(Ljudpost)	5	0	100%
Särskrivning	5	0	100%
Versal(Egennamn)	4	0	100%
Kongruens(Kongruensfel)	4	0	100%
Variation(Sambandsord)	4	0	100%
Skiljetecken(Frågetecken)	2	2	0%
Variation(Adjektiv)	2	0	100%
Kongruens(Predikat)	1	0	100%
Variation(Verb)	0	0	N/A
Skiljetecken(Utropstecken)	0	0	N/A
Variation(Meningsbörjan)	0	0	N/A
Pronomen(Object/Subject)	0	0	N/A
Ordföljd(I bisats)	0	0	N/A
Ordföljd(Upprepade ord)	0	0	N/A
Komparation(Perifrastisk)	0	0	N/A
Komparation(Suffixkomparation)	0	0	N/A
Ordföljd(I indirekt fråga)	0	0	N/A
Granska/Service total	173	57	67%

Table 7: Error analysis from 4 random texts.

5 Conclusions and recommendations

5.1 Meeting the requirements from Skolverket

In section 2.2, we investigated the requirements needed to evaluate the writing aspects of students, set by Skolverket, with the use of NLP. It is clear that no single NLP technology can cover all the aspects and that for some aspects, such as taking into account the intentions of the writer, there is no technology present to cover it at all. Nonetheless, by combining several technologies a vast majority of the aspects can to some extent be covered.

A major obstacle in order to evaluate some of the aspects is the need for large and relevent corpuses. If provided however, some interesting solutions are possible. As school essays are written based on different genres, and Skolverket has provided criteria to be met for each genre, it would be interesting to see the results of a machine learning algorithm that classifies student texts according to genres.

I think the struggle for such a solution would be how go give relevant feedback to the student. Such a classifier would most likely have different values, such as appropiate vocabulary, it can be difficult to convey to the student how it should adjust the text based on these values. In most cases it can only convey that its vocabulary is not appropriate enough and can only give general instructions on how to improve it, such as using certain words more often. If however it could instead pinpoint weak points in the text and feedback relevant suggestions then it would be a major attribute to students.

5.2 Adjusting Granska

Granska has been adjusted in several ways in the course of this project. It is now possible for anyone to contribute to the Granska project as it is available as an open-source project on GitHub. As it is also now possible to install Granska as a dynamic library, with an API separating its loading features from its scrutinizing features, it is now more adapted to handle multiple queries as part of a web service. Compatible with both Linux and Solaris, one can more freeely use the operative system of ones choosing.

A number of new features have been added to Granska as well. For starters, all the rules triggered during scrutinization are saved in a table and are outputed in a structured XML format along with the rest of the Granska output. The rule library has been extended with new categories and rules specially designed to cover aspects associated with writing criteria from Skolverket and with that also came extensions to the Granska rule language and the Stava API. Finally a test program was created to make sure the dynamic library is installed and works properly.

5.3 Analysis of the result

The result shows that the tool works and that relevant feedback based on aspects from Skolverket is outputed. It is verbose, but only about half as verbose as the teacher. A few aspects find few, if any, errors among the texts while another few finds the majority of all errors found. The precision on *Verb(Tempusbyte)* (68%), for example, is low while others score a perfect 100%.

It is easy to jump to the conclusion that *Verb(Tempusbyte)* is not good enough and should be removed or that some of the aspects who do not find any errors are unuseful. What the result is not displaying however is the feedback from teachers and students using the tool. It was a goal to include that part in this thesis but no such evaluations were ever made. Luckily for this project however, the evaluations are going to made by the continuing work of Ingivos AB who by now have their own prototype which, among other technologies, is running all the technologies from this project. And as Granska now is an open-source project, hopefully more contributions will be made to it in the future.

References

[1] http://www.csc.kth.se/tcs/projects/granska/rapporter/rulelang20001121.pdf, November 200.

[2] http://office.microsoft.com/en-us/word/, April 2014.

[3] https://www.openoffice.org/product/index.html, April 2014.

[4] http://www.jspell.com/public-spell-checker.html, April 2014.

[5] https://github.com/viggokann/granska, April 2014.

[6] Utarbetad av Svenska Språknämnden. Språkriktighetsboken. 2005.

[7] Utgivna av Svenska Språknämnden. Svenska skrivregler. 1998.

[8] Johnny Bigert, Viggo Kann, Ola Knutsson, and Jonas Sjöbergh. Grammar checking for swedish second language learners. In *Chapter in CALL for the Nordic Languages, Copenhagen Studies in Language*, volume 30, pages 33–47. Samfundslitteratur, 2005.

[9] Johan Carlberger, Rickard Domeij, Viggo Kann, and Ola Knutsson. A swedish grammar checker. 2002.

[10] Johan Carlberger and Viggo Kann. Implementing an efficient part-of-speech tagger. 1999.

[11] Rickard Domeij, Joachim Hollman, and Viggo Kann. Detection of spelling errors in swedish not using a word list en clair. *J. Quantitative Linguistics*, 1:1–195, 1994.

[12] Barbara J. Grosz, Scott Weinstein, and Aravind K. Joshi. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21:203–225, 1995.

[13] E. Miltsakaki and K. Kukich. Evaluation of text coherence for electronic essay scoring systems. In *Nat. Lang. Eng*, volume 10, pages 25–55. Cambridge University Press, March 2004.

[14] Robert Östling, André Smolentzov, Björn Tyrefors Hinnerich, and Erik Höglin. Automated essay scoring for swedish. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–47, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[15] Skolverket. Språket på väg. Edita Västra Aros, 2011.

[16] Erik Andersson Ulf Teleman, Staffan Hellberg. Svenska akademiens grammatik. Svenska Akademien, 1999.

[17] Östen Dahl. Grammatik. 2003.

Appendix A

Skriva – lärarmatris

	Förnivå	Steg 1	Steg 2	Steg 3	Steg 4
Mottagare *Hur texten fungerar för den som ska läsa den* *Kommunikativa och kognitiva funktioner*		Texten är uppbyggd så att läsaren i stort sett kan ta till sig och förstå innehållet.	Texten är begriplig för läsaren som också förstår textens syfte (till exempel roa, övertyga, informera …)	Texten står på egna ben och kommunicerar ganska väl med läsaren i enlighet med textens syfte (till exempel roa, övertyga, informera …)	Texten står på egna ben och kommunicerar mycket väl med läsaren i enlighet med textens syfte (till exempel roa, övertyga, informera …)
			Eleven har oftast gjort lämpliga språk- och stilval utifrån textens syfte och tänkta läsare.	Eleven har gjort lämpliga språk- och stilval utifrån textens syfte och tänkta läsare.	Eleven har gjort lämpliga och finessrika språk- och stilval utifrån textens syfte och tänkta läsare.
				Interaktionen med läsaren främjas av de språkliga valen.	Interaktionen med läsaren främjas av de lyckade språkliga valen.
			Texten är logiskt uppbyggd.	Texten är relativt väl sammanhållen och har få omotiverade tankeluckor.	Texten är väl sammanhållen, väl avgränsad och har inga omotiverade tankeluckor.
					Medvetna och motiverade tomrum förekommer.
		Texten är bunden till källan (när skribenten använder källor).	Texten visar på viss självständighet gentemot källan (när skribenten använder källor).	Förhåller sig självständigt gentemot källan (när skribenten använder källor)	Låter olika källor samspela (när skribenten använder källor).
Textens genre (läs mer på sidan 12–13)		Försöker att följa genrens mönster för struktur och språk.	Följer något mekaniskt genrens mönster för struktur och språk.	Följer genrens mönster för struktur och språk på ett levande sätt.	Följer genrens mönster för struktur och språk på ett levande sätt och har ibland egna kreativa lösningar. Innehåll, struktur och språk samspelar på ett bra sätt.
Helhetsstruktur och språk – globala och lokala nivåer (läs mer i del I på sidan 7–8) *Textnivå*		Textens disposition fungerar delvis. Rubriker finns och har anknytning till innehållet	Textens disposition är mer genomtänkt och fungerar genomgående ganska väl. Rubrikerna täcker textens innehåll.	Textens disposition är tydlig, väl genomtänkt och fungerar väl. Rubrikerna täcker textens innehåll och fungerar	Textens disposition är avancerad och samtidigt tydlig och mycket väl genomtänkt. Rubrikerna är tydliga, täcker textens innehåll och fungerar utmärkt.

	Förnivå	Steg 1	Steg 2	Steg 3	Steg 4
Styckenivå		Styckeindelning förekommer.	Styckeindelningen fungerar genom att varje stycke är en sammanhållen helhet.	Väl fungerande styckeindelning, ofta med tydliga kärnmeningar.	Väl fungerande styckindelning med tydliga kärnmeningar där det är lämpligt. Variation på styckenas längd inspirerar läsningen och markerar deras olika tyngd.
		Bindningar förekommer.	Bindningar förekommer och fungerar ganska väl.	Meningarna är bundna genom olika sorters bindningar, vilka fungerar väl.	Bindningar inom och mellan meningar är varierade, väl valda och fungerar utmärkt.
Meningsnivå		Meningarna har oftast en fungerande syntax och är: · enkelt konstruerade (inte så många bisatser) · ganska korta · inledda med subjektet.	Meningarna är mer varierade men fortfarande: · ofta enkelt konstruerade (inte så många bisatser) · ofta ganska korta · ofta inledda med subjektet.	Det finns i sättet att konstruera meningar en viss syntaktisk variation, dvs. en viss blandning av: · enkel och mer komplicerad syntax · långa och korta meningar · varierande satsdelar i fundamenten · ofta konstruerade med lätta fundament, dvs. högertunga.	Meningarna har en syntaktisk variation som fungerar mycket väl utifrån syfte, tänkta läsare och sammanhang. Det finns: · enkel och mer komplicerad syntax · långa och korta meningar · varierande satsdelar i fundamenten · ofta konstruerade med lätta fundament, dvs. högertunga. Den syntaktiska variationen ger ett behagligt flyt i texten.
		Satsradningar kan förekomma.	Satsradningar kan förekomma och då är satserna ofta skilda åt med kommatecken.	Satsradningar är sällsynta. Satserna är ofta kopplade med sammanhangssignaler och konjunktioner.	Om satsradningar förekommer är de stilistiskt medvetet genomförda.

	Förnivå	Steg 1	Steg 2	Steg 3	Steg 4
Uttrycks- och ordnivå		Uttryck och ord fungerar ibland mindre bra. Orden kan oftast karaktäriseras som vardagliga, allmänna, konkreta och något talspråkliga. Ordvalet är inte alltid så varierat och kanske inte alltid passar för textens syfte och mottagare.	Uttryck och ord fungerar ganska väl i sina sammanhang. Många ord kan karaktäriseras som vardagliga, vanliga, allmänna, konkreta och talspråkliga. Texten visar viss variation i ordval.	Uttryck och ord fungerar oftast väl i sina sammanhang. Det finns en ansats till att använda ord som kan karaktäriseras som mindre vardagliga, specifika, abstrakta och skriftspråkliga. Texten visar variation i ordval.	Uttryck och ord är mycket väl valda och fungerar mycket väl i sina sammanhang. Det kan finnas exempel på nyskapande vad gäller uttryck och ord, vilket berikar läsningen. Texten visar variation och omsorg i ordval.
Normer och språkriktighet		Elever på denna nivå bryter sällan mot normer i stilistiskt syfte.	→	→	Elever på denna nivå kan ibland medvetet bryta mot normer i stilistiskt syfte.
Styckemarkering		Markerar på eget sätt att nytt stycke börjar, till exempel genom s.k. hybridstycke.	Blandar olika markeringar för nytt stycke: hybridstycke, blankrad och indrag, utan att man som läsare förstår systemet.	Markerar nytt stycke med blankrad eller indrag.	Markerar nytt stycke med blankrad eller indrag. Kan om det är lämpligt markera nytt stycke med blankrad och indrag, för att skilja på större avsnitt och stycken. Som läsare förstår man systemet.
Meningsbyggnad		Oftast väl fungerande vid enklare meningsbyggnad, dvs. vid huvudsatser utan bisatser. Undviker en mer komplicerad syntax eller lyckas inte konstruera mer komplexa meningar väl fungarande.	Prövar en mer komplex meningsbyggnad och lyckas då ibland väl med konstruktionen.	Oftast väl fungerande meningsbyggnad även i långa och mer komplexa konstruktioner.	Väl fungerande eller i stort sett väl fungerande meningsbyggnad i såväl enklare som mer komplexa konstruktioner.

	Förnivå	Steg 1	Steg 2	Steg 3	Steg 4
Konstruktion av uttryck		Använder enkla och vanliga konstruktioner som oftast är väl fungerande. Undviker att bygga ut fraser och att använda mer komplicerade uttryck eller lyckas inte konstruera mer komplexa fraser och uttryck korrekt.	Prövar mer komplicerade fraser och uttryck och lyckas då ibland väl med konstruktionen. Har vissa svårigheter t.ex. med val av prepositioner.	Oftast väl fungerande konstruktion av fraser och uttryck även i långa, ovanliga, mer komplexa och skriftspråkliga konstruktioner.	Väl fungerande eller i stort sett korrekta konstruktioner av både enkla och mer komplicerade fraser och uttryck.
Användning av ord		Använder de flesta ord på ett lämpligt sätt. Undviker ord som kan karaktäriseras som svåra eller ovanliga.	Prövar ibland ord som kan karaktäriseras som svåra eller ovanliga och lyckas då ganska väl med användningen.	Använder både enklare och ord som kan karaktäriseras som svåra eller ovanliga och då oftast på ett lämpligt sätt för sammanhanget.	Använder en variation av ord på ett kreativt och lämpligt sätt.
Stavning		Stavar de flesta ord på ett för genren lämpligt sätt Gör vissa systematiska stavfel.	Prövar ibland ord som kan karaktäriseras som svåra eller ovanliga och lyckas då ganska väl med stavningen.	Stavar oftast på ett för genren lämpligt sätt.	→
Användning av tecken		Använder oftast punkt och i förekommande fall frågetecken och utropstecken på ett lämpligt sätt. Punkt, frågetecken och utropstecken följs av versal.	Använder punkt, frågetecken, utropstecken korrekt och kommatecken på ett lämpligt sätt så att läsningen av texten underlättas. Använder oftast versaler och gemener på ett lämpligt sätt.	Använder en variation av skiljetecken och andra skrivtecken och då oftast på ett lämpligt sätt. Använder versaler och gemener på ett lämpligt sätt	Använder en variation av skiljetecken och andra skrivtecken på ett lämpligt och kreativt sätt.
Texten i sin helhet		Visar att eleven är på rätt väg i sin skrivutveckling. Eleven behöver tydligt stöd för att utveckla innehåll, struktur och språk i sin text.	→	→	Visar att eleven kommit långt i sin skrivutveckling. Innehåll, struktur och språk samspelar väl.

Appendix B

Genrebeskrivningar

Så här kan man kortfattat beskriva genrerna berättelse, insändare och brev.

Berättelse

- Fokus på händelser; har en handling och därmed tid i sig
- Reflexioner, inre monologer
- Miljöbeskrivningar – gestaltningar; har en riktning, detaljer och helhet (zooma in och zooma ut)
- Personbeskrivningar – gestaltningar
- Tonvikt på att berätta och beskriva
- Dialog (främst i skrivna berättelser)
- Ofta finns tomrum i texten
- Budskapet bildar en röd tråd genom berättelsen
- Grundläggande struktur: ”början – mitt – slut”; ofta en spänningskurva som kan tecknas ungefär som ett ”W”: allt är bra eller som vanligt – problem uppstår – lösning verkar nära – det lyckas inte – slutet gott eller ljusning anas
- Kronologi eller annan mer komplicerad dispositionsprincip
- Genomgående tempus: presens (som hänger ihop med perfekt och futurum) eller preteritum (som hänger ihop med pluskvamperfekt och futurum preteritum)
- Rörliga eller dynamiska verb är ofta viktiga
- Sammanhangssignaler som uttrycker tid är vanliga: *medan, samtidigt, efter ett tag, därefter, slutligen*

Insändare

- Innehåller tes eller huvudtanke och argument
- Argumentens kvalitet – relevans, hållbarhet (sakargument – sant; värdeargument – godtagbart)
- Nyansering (motargument – bemöta, visa förståelse för; inte fula knep, personangrepp eller vantolkningar)
- Den egna rösten, engagemang och känslor
- Grundläggande struktur: inledning – väcka uppmärksamhet och leda in i ämnet; tes eller huvudtanke; argument för argument med belysande och förstärkande moment; avslutning där huvudtanken upprepas
- Logik i resonemang och disposition; analyserande; röd tråd; bindningar
- Genomgående tempus: presens (som hänger ihop med perfekt och futurum)
- Sammanhangssignaler som uttrycker motsättning, slutsats eller emfas: *fördelarna, nackdelarna, alltså, slutsatsen blir, tydligen, i själva verket, å ena sidan ... å andra sidan, visserligen ... men så är, även om ... så är, för det första ... för det andra, framför allt, inte minst*

Brev med informativ och förklarande text

- Relevant information bearbetad och sammanfogad med författarens tidigare kunskaper och erfarenheter
- Logisk text där tankar förklaras; syftet med texten framträder
- Resonerar, analyserar och kommenterar
- Skriver så att viktiga stycken, meningar och ord framträder för läsaren; röd tråd; bindningar; kärnan (centrala tankar, begrepp och teman) i texten och kärnmeningar (meningar som sammanfattar respektive stycke) i styckena tydliggörs; meningsbärande innehållsord framträder
- Dispositionsprincip, till exempel:
  - tid
  - orsak
  - tematisering

  Genomgående tempus: presens (som hänger ihop med perfekt och futurum) eller preteritum (som hänger ihop med pluskvamperfekt och futurum)
- Sammanhangssignaler som uttrycker orsaksförhållanden och som exemplifierar och och preciserar är vanliga: *på grund av detta, därför, men, som en följd av, anledningen till, till exempel, i synnerhet, framför allt, å ena sidan…å andra sidan*
- Genremönster: hälsningsfraser i inledning och avslutning samt ort och datum