The Uppsala Corpus of Student Writings
Corpus Creation, Annotation, and Analysis

Beáta Megyesi[1], Jesper Näsman[2], Anne Palmér[3]
Department of Linguistics and Philology[1,2], Department of Scandinavian Languages[3]
Uppsala University, Sweden
beata.megyesi@lingfil.uu.se, jesper.nasman@lingfil.uu.se, anne.palmer@nordiska.uu.se

Abstract

The Uppsala Corpus of Student Writings consists of Swedish texts produced as part of a national test of students ranging in age from nine (in year three of primary school) to nineteen (the last year of upper secondary school) who are studying either Swedish or Swedish as a second language. National tests have been collected since 1996. The corpus currently consists of 2,500 texts containing over 1.5 million tokens. Parts of the texts have been annotated on several linguistic levels using existing state-of-the-art natural language processing tools. In order to make the corpus easy to interpret for scholars in the humanities, we chose the CoNLL format instead of an XML-based representation. Since spelling and grammatical errors are common in student writings, the texts are automatically corrected while keeping the original tokens in the corpus. Each token is annotated with part-of-speech and morphological features as well as syntactic structure. The main purpose of the corpus is to facilitate the systematic and quantitative empirical study of the writings of various student groups based on gender, geographic area, age, grade awarded or a combination of these, synchronically or diachronically. The intention is for this to be a monitor corpus, currently under development.

Keywords: student writings, digital humanities, educational applications

1. Introduction

Tools developed in computational linguistics for the automatic analysis of texts can be useful in many subject areas in the humanities and social sciences, enabling more consistent, large-scale quantitative linguistic analyses of various kinds of texts. This new field of digital humanities may introduce new approaches to research in the different disciplines. However, many natural language processing tools cannot be used directly by scholars since they often require programming skills and some knowledge of computational linguistics. Digital humanities can help researchers in the humanities take advantage of the resources and tools available in language technology without requiring them to have the technical expertise normally associated with these kinds of tools. This paper describes such an effort, a collaboration between scholars in the Swedish language and computational linguists to perform a quantitative analysis of Swedish student writings over time based on automatic linguistic analysis. The resource we present is unique not only to Swedish scholars examining student writings and their development over time, but also to the entire academic community since the set-up can generally be applied to other languages.

The Research Group for National Tests in Swedish and Swedish as a Second Language at Uppsala University designs national tests for students in elementary school years 3, 6, 9 and in upper secondary school on behalf of the Swedish National Agency for Education. These tests are mandatory and are aimed at providing equivalent and fair grades to every student[1]. The research group regularly receives a subset of the completed tests in order to carry out follow-up studies. Digitization of these tests brings new opportunities for research on student writing over time. Calculating linguistic features such as part-of-speech distribution, syntactic structure or average word length of texts can be done easily within a few seconds. Researchers interested in investigating student writing are keen to explore the possibilities offered by these new technologies.

The focus of this study is on building a corpus of student essays, the Uppsala Corpus of Student Writings, and annotating them on several linguistic levels by adapting existing state-of-the-art natural language processing tools developed for standard language. In addition, we present an application for large-scale linguistic analyses of these texts, allowing a user-friendly search and information extraction function, which enables scholars with few or no programming skills to easily explore student writings over time based on the analysis provided by the tools.

In Section 2., we give a brief summary of the research carried out on student writings from the perspective of Swedish studies in the humanities to give an idea of the research questions and opportunities that computational linguistics could offer in terms of large-scale quantitative studies. Following this introductory section, we present the corpus data and the corpus format in Section 3., and 4. We then give an overview of the automatic linguistic annotation tools used and an evaluation of the components in Section 5. In Section 6., we describe some of the linguistic characteristics of student essays based on our corpus data. Lastly, in Section 7., we conclude the paper and indentify some future challenges.

2. Work on Student Writings

Research on student writings can roughly be described as focusing on text analysis, or on text in context. In Scandinavia, studies on student writings with an emphasis on quantitative measures were frequently carried out in the late 20th century, but have also been conducted in the 21st cen-

[1] http://www.skolverket.se/bedomning/nationella-prov

tury, predominantly by combining quantitative and qualitative methods. Moreover, a large volume of research focusing on text in context has been published since the turn of the millennium. The most important study in the quantitative tradition is by Hultman and Westman (1977), which compares 151 student essays from a national test with adult texts. Statistical information about vocabulary, distribution of parts-of-speech, syntax and language errors in the essays are correlated with grades awarded and show that essays awarded high grades can be characterized by a varied vocabulary, nominal style and few errors. Larsson (1984) continued this quantitative tradition in his examination of the notion of language ability, by correlating students' texts and their grades with tests on vocabulary knowledge and reading, as well as with extra-linguistic factors. Quantitative methods combined with qualitative methods were used in a study of the paragraphing in student texts (Strömquist, 1987) and student writing strategies in school essays (Garme, 1988). Later examples of research using quantitative measures include e.g. Östlund-Stjärnegårdh (2002) about assessing student writings on the borderline between pass or fail, and Nordenfors (2011), who focused on writing development over time in compulsory schools. Magnusson and Johansson (2009) examine texts written by students of Swedish as a first and second language and argue that measures like nominal ratio, word/lexical variation index and word length, all used by Hultman and Westman (1977), can be used as indicators of text quality.

Research in the quantitative tradition is interesting given its ability to describe the main characteristics of student writings. However, one problem is that the samples used are often rather small, which causes severe bias. Another problem is that most studies only use a few measures; among the studies mentioned above, Hultman and Westman (1977) use the most diversified methods. Access to more important samples of student texts as well as a tool for automatic analysis that does not require extensive technological know-how would greatly facilitate future research.

3. Corpus Data

The proprietor of the essays is the Research Group for National Tests in Swedish and Swedish as a Second Language, part of the Department of Scandinavian Languages, Uppsala University. This group, appointed by the Swedish National Agency for Education, is responsible for developing and designing the national tests in Swedish as a school subject. For research and follow-up studies, the group has established an archive with student answers to the tests as well as a database with test results. The collection of student essays in this archive is now extensive and currently contains 80,000 essays.

The essays represent national tests in two school subjects: Swedish and Swedish as a second language. The essays were written by students from the age of nine in year three of primary school to the age of nineteen in upper secondary school, who are studying either Swedish or Swedish as a second language. The tests have been collected since 1996, and some have been digitized. The tests have been given under two different sets of national curricula, the 1994 curricula (Lpo94 and Lpf94), in effect from 1994 to 2010, and the 2011 curricula (Lgr11 and Lgy11), in use since 2011.

The corpus currently contains 2,500 texts consisting of more than 1.5 million tokens. Table 1 summarizes the corpus data available with information about the school level (years 3, 5, 6, and 9 of compulsory school, and the first and last year of upper secondary school), the age of the student, the type of school and curricula, the number of essays, the number of tokens from each subset, and the average number of tokens per essay.

As Table 1 shows, the essays are not evenly distributed across the age groups. The reason why the majority of the essays were taken from the final years of upper secondary school is that there is currently a knowledge gap in Sweden concerning the level of student writing competence when students finish school. Our intention is to help fill this knowledge gap, and the large data set of essays from upper secondary school will serve this purpose well. In addition, we also wanted to take advantage of one corpus that had already been produced at the Research Group for National Tests - but not annotated - as a starting point. This corpus contains essays from national tests produced by the group since 1996, tests for both compulsory school (through lower secondary school) and upper secondary school. It constitutes an interesting overview of student essays from different tests, text genres and years. Whereas the essays from upper secondary school already constitute what we would consider as rich data, the selection of essays from compulsory school needs to be further developed. That will be the next step.

3.1. Preparation of the Essays

Since the archive is not digital but in paper format, substantial manual efforts are needed to prepare the essays for annotation. The majority of essays are handwritten, but nowadays schools may allow their students to write essays on computers so there is a growing number of essays in printed format, although still submitted on paper. The various steps of essay preparation, which differ slightly depending on the type of script (handwritten or printed), are shown in Table 2. Handwritten essays are transcribed by a human annotator, while printed essays written on a computer are scanned as PDF files, converted into text files, manually validated and corrected if necessary. Each essay is then coded with metadata information, proofread and edited if necessary.

The preparation is carried out by staff from the research group and by students from the teacher training program. The process is quite time-consuming, especially when handwritten essays are involved. The estimated processing time required is 5 minutes per 100 words for handwritten essays and 3 minutes for printed essays. There is great variation in the time it takes to process the essays depending on the quality of the handwriting, the efficiency of the staff, etc.

This process is intended to be optimized in the future through the use of better image processing techniques for handwritten essays, and through the development of a web-based portal where students can upload their essays digitally in computer-readable format.

Level	Age	School level and curriculum	Number of essays	Number of tokens	Tokens per essay
C-3	9	Compulsory, Lpf94 + Lgy11	91	8,644	95
C-5	11	Compulsory, Lpf94	66	13,121	199
C-6	12	Compulsory, Lgr11	47	17,741	377
C-9	15	Compulsory, Lgr94 + Lgr11	249	137,689	553
US-1	16	Upper Secondary, Lgy11	131	76,521	584
US-3	18	Upper Secondary, Lgy11	410	347,836	848
GY-3	18	Upper Secondary, Lpf94	1,506	1,055,468	701
Total			2,500	1,657,020	663

Table 1: Distribution of the subset of texts by school year, given as number of texts, sentences and tokens, and average number of tokens per essay used in the pilot study.

Handwritten essays	Printed essays
Transcription	Scanning-conversion-editing
Coding	Coding
Proofreading and final editing	Proofreading and final editing

Table 2: Preparation of the essays.

METADATA	Description
TEXT-ID	each text receives a letter and a number
TEST	test type: GY, KP3, KP1, 9, 6, 5, 3
DATE	year and when available, semester in which the essay was written
GENRE	text genre: argumentative (ARG), explanatory (UTR), narrative (BER), instructional (INST), descriptive (BESK)
GRADE	grade awarded to the text - the scale varies depending on the year the test was produced e.g. A-B-C-D-E-F; IG-G-VG-MVG; EN-G-VG-MVG
GENDER	gender of the student: male (M), female (K)
SUBJECT	Swedish (Sv) or Swedish as a second language (SVAS)
PERMISSION	permission of student to publish the essay (T) or not (ET)
AREA	name of the municipality the school is in, e.g. Uppsala, Stockholm
EDUCATION	program of studies in upper secondary school, e.g. BF, BA, EE, ES, FT, HA, HV, HT, HU, IN, NB, ...
FORMAT	how the essay was produced: handwritten (H) or typed on a computer and printed (D)

Table 3: Metadata information.

3.2. Coding the Essays

Each essay is given a code indicating metadata information about 11 aspects of the essays. Information is available about the year and semester when the texts were produced, the genre the students were asked to produce, the grade awarded, the gender of the student, the type of Swedish the student was studying (Swedish or Swedish as a second language), and other data. The full list of metadata information encoded is shown in Table 3.

The various metadata fields are combined and given in brackets in the beginning of each text as <TEXT-ID TEST DATE GENRE GRADE GENDER SUBJECT PERMISSION AREA EDUCATION FORMAT>. For example, <C14 KP1 VT13 UTR A M SV T UPPSALA NV D> encodes text C14 from the selection KP1, produced during the spring of 2013 (VT13) which is in the explanatory genre (UTR) and was awarded an A, written by a male (M) who studied Swedish (SV). He gave his permission for the text to be used for research (T), his school is located in Uppsala, and he was in the scientific studies program (NV). The text was typed on a computer (D) when he wrote the essay.

The metadata information is designed to enable filtration of the essays by different groupings based on the research questions. The researcher may, for example, compile a subgroup of essays from the KP3 test type for every year the test was given, extract explanatory texts awarded an A written by students of both genders in the school subject Swedish in all municipalities and all upper secondary school programs and consider only printed essays written on computers. This subgroup may then be compared with a similar subgroup but with essays awarded a C, etc.

Some metadata information may be missing for some tests given that the eassays have been collected over a lengthy period; some essays were already digitized before this project started so some information may have been lost.

3.3. Development Set

We selected a subset of data for a pilot study to test whether the proposed corpus format and the automatic linguistic analysis and annotation would be appropriate for scholars

studying these types of student writings. The pilot subset consists of 245 completed tests, which were sampled to represent different levels of achievement with respect to age, school year, gender and text type, as illustrated in Table 4. The pilot data set was carefully selected to include typical examples of student writings representing various grades awarded from low grades to high ones for different age groups.

Level	Texts	Sentences	Tokens
3	24	270	3,390
5	16	337	5,348
6	23	851	12,233
9	78	2,627	47,264
Upper Sec. School	104	4,239	89,698
Total	245	8,234	157,933

Table 4: Distribution of the development set by grade awarded given as number of texts, sentences and tokens used in the pilot study.

4. Format

In order to make the corpus easy to interpret for scholars in the humanities, that is, to provide a format that is easy to read and understand, we use the CoNLL-X[2] shared task format instead of an XML-based representation. In particular, since we are interested in the morpho-syntactic annotation, we have used the CoNLL-U[3] representation developed for the universal dependency annotation with some minor adaptations in order to represent corrections of student writings as spelling and grammatical errors, which occur frequently in these types of texts.

All annotations are encoded in plain text files in UTF-8. Sentences consist of one or more lines of words where each line represents a single word/token with a series of 11 fields with separate tabs for various annotation types. New sentences are preceded by a blank line, which marks sentence boundaries. Comment lines starting with hash (#) are also allowed and may be used for metadata information (such as sentence numbering) for the sentence following immediately. The word lines contain information about the text index number to enumerate the paragraph and sentence in which the token occurs, the index number of the token in the sentence, the word form as it appears in the original text, the normalized form as given by the spelling correction, lemma, part-of-speech and morphological information, and the syntactic annotation represented as dependency structures. Empty values are marked as underlines (_). Table 5 describes the fields that represent the analysis of each token. Table 6 exemplifies the annotation when all tokens are correctly spelled by the student. When a word is misspelled or a compound is written incorrectly as two words, the correct spelling is given in the fourth column. In the case of misspelled compounds written as two or more words instead of one, the corrected version is given first with the merged index number of the tokens included followed by the linguistic analysis. The original, incorrectly written words are shown in the lines following immediately with the original index numbers of the token without any linguistic analysis. An example of words written incorrectly, or misspelled, is shown in Table 7[4] for the word *inspirationskälla* (inspiration source) which was wrongly spelled as two words: *inspirations källa*, and the word *texten* (the text) incorrectly spelled as *teksten*.

5. Automatic Annotation

In order to automatically process and annotate the texts, we use state-of-the-art natural language processing tools trained on Swedish standard texts. The annotation process is illustrated in Figure 1. Each module is described in the following subsections.

5.1. Preprocessing

First, each digitized essay is preprocessed by restructuring the meta-textual information from the original files. This step is carried out manually to find any inconsistent manual mark-up of the information about the texts, by searching for and checking each entry of metadata. Furthermore, each text, if necessary, is automatically converted from Microsoft Word or any other format into a text file (UTF-8) prior to automatic annotation.

5.2. Tokenization

After preprocessing, tokenization is used to separate the tokens and segment the sentences. We tested two different state-of-the-art tokenizers for Swedish: a built-in tokenizer of the PoS tagger Stagger (Östling, 2013), and the tokenizer Svannotate, which was developed to automatically process texts based on the Swedish Treebank data (Nivre et al., 2008). The output of the tokenizers was compared, and a total of 10 differences were found. Only one of these differences was an error made by Svannotate, while 9 were errors made by Stagger. We chose the Svannotate rule-based tokenizer and sentence segmenter in the automatic processing. When evaluating the tokenizer on student writings, errors that occurred are due in part to the inconsistent use of punctuation marks – that is, when a sentence does not always end with an appropriate punctuation mark, either because abbreviations are not always spelled correctly, or a new sentence does not always begin with a capital letter.

Since the annotation pipeline is modular, the user has the option of tokenizing a text and manually correcting it, and then using the corrected version for the remaining steps (normalization, PoS tagging and syntactic annotation).

5.3. Normalization

After the texts have been tokenized and the sentences segmented, we apply spelling correction to detect and correct possible spelling errors. We use HistNorm (Pettersson et al., 2013), originally developed for the automatic correction of historical word forms with a large variation in possible

[2]CoNLL-X shared task representation: http://ilk.uvt.nl/conll/

[3]CoNLL-U representation format for Universal Dependencies: http://universaldependencies.github.io/docs/format.html

[4]Please note that due to lack of space, we do not show all columns, we have left out the fields TEXT ID and UPOS.

FEATURE	Description
TEXT ID	Text-Paragraph-Sentence index, integer starting at 1 for each new text, paragraph and sentence
TOKEN ID	Token index, integer starting at 1 for each new sentence; may be a range for tokens with multiple words
FORM	Word form or punctuation symbol
NORM	Corrected/normalized token (e.g. in case of spelling error)
LEMMA	Lemma or stem of word form
UPOS	Part-of-speech tag based on universal part-of-speech tag
XPOS	Part-of-speech tag based on the Stockholm-Umeå Corpus; underscore if not available
FEATS	List of morphological features; underscore if not available
HEAD	Head of the current token, which is either a value of ID or zero (0)
DEP	Dependency relation to the HEAD (root iff HEAD = 0) based on the Swedish Treebank annotation
DEPS	List of secondary dependencies (head-deprel pairs)
MISC	Any other annotation

Table 5: Annotation representation for each token and field.

TEXT ID	ID	FORM	NORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEP	ENGLISH
2-3	1	Ödlor	Ödlor	ödla	NOUN	NN	UTR |PLU |IND |NOM	2	SS	Lizards
2-3	2	gillar	gillar	gilla	VERB	VB	PRS |AKT	0	ROOT	like
2-3	3	att	att	att	IE	IE	_	2	OO	to
2-3	4	äta	äta	äta	VERB	VB	INF |AKT	3	IF	eat
2-3	5	insekter	insekter	insekt	NOUN	NN	UTR |PLU |IND |NOM	4	OO	insects

Table 6: Example of the extended CoNLL-U shared task format for the sentence *Lizards like to eat insects.*

ID	FORM	NORM	LEMMA	XPOS	FEATS	HEAD	DEP	ENGLISH
1	Min	Min	min	PS	UTR |SIN |DEF	2-3	DT	My
2-3	inspirationskälla	inspirationskälla	inspirationskälla	NN	UTR |SIN |IND |NOM	4	SS	inspiration-source
2	inspirations							
3	källa							
4	kommer	kommer	komma	VB	PRS |AKT	0	ROOT	comes
5	från	från	från	PP	_	4	RA	from
6	teksten	texten	text	NN	UTR |SIN |DEF |NOM	5	PA	the text

Table 7: Example of the extended CoNLL-U shared task format for the sentence *My source of inspiration comes from the text.* with two misspelled words, the compound *inspirationskälla* and the word *text*.

Figure 1: The process of automatic annotation.

spellings to their modern variant. There are currently two normalization methods implemented: Levenshtein-based normalization and normalization based on statistical machine translation (SMT). For Swedish, results show that character-based SMT used for spelling correction gives the highest accuracy, 92.9% when applied to historical data (Pettersson et al., 2014). Further study is needed to adapt this normalization tool to student writings for higher accuracy.

In addition to spelling errors, erroneously split compounds that should be written as a single word occur frequently in student writings in Swedish. A further problem is that split compounds sometimes also change the meaning of the expression, which could cause ambiguity, such as in *kassamedarbetare*, which is translated into *cashier*, vs. *kassa medarbetare*, which is translated into *useless employees*. The method for identifying split compounds is currently under development. Usually, compound detection can be taken care of by a rule-based system, such as that presented by Öhrman (1998) but it often requires PoS-tagged tokens as input. Here, any sentence where a split compound is identified thus needs the insertion of new lines, which makes any linguistic analysis prior to compound detection problematic.

5.4. Morpho-Syntactic Annotation

For the PoS and morphological annotation of the normalized texts, two commonly used PoS taggers for Swedish,

HunPos (Halácsy et al., 2007) and Stagger (Östling, 2013), were evaluated and compared after being applied to the test data. The taggers were trained on the second version of the Stockholm Umeå Corpus (Gustafson-Capková and Hartmann, 2006), which is the standard reference corpus for Swedish. In earlier studies, HunPos has achieved 95.9% accuracy (Megyesi, 2008) and Stagger 96.6% (Östling, 2013) when trained on SUC 2.0. Since a correctly annotated gold standard of student writings does not exist, the differences between the outputs of the two taggers were taken into consideration when errors introduced by both taggers were considered correct. The results show that Stagger performed slightly better than HunPos using both plain PoS tags and PoS tags with morphological information. Stagger also includes a lemmatizer which is used in connection with PoS tagging.

Stagger was recently re-implemented, and released under the name Efficient Sequence Labeler (efselab)[5]. Since efselab processes text significantly faster and also has support for universal dependencies for Swedish (Nivre, 2014) which could also be selected for the syntactic analysis, we chose efselab as the default tagger.

5.5. Syntactic Annotation

As the last step in the linguistic annotation process, the syntactic analysis is carried out in the form of a dependency structure. Currently, we use SweMalt, the Swedish MaltParser model, which is a single malt configuration for parsing Swedish text with MaltParser (Nivre et al., 2006), version 1.7.2. The parser was trained on the Swedish Treebank (Nivre et al., 2008) and the SUC PoS tagset with morphological features. Consequently, during parsing, we use PoS annotation based on the SUC tagset as input to the parser.

Moreover, a recently developed model based on universal dependencies developed for Swedish (Nivre, 2014) is underway in the Swedish annotation pipeline of efselab and has been implemented as a possible choice for the syntactic annotation of the corpus. The Swedish annotation pipeline of efselab is adapted from the Swedish Treebank pipeline and contains a tokenizer, a PoS tagger using efselab with a SUC model, which also converts the SUC tagset into the Universal PoS tagset (including differentiation between auxiliary verbs and other verbs), lemmatization and dependency parsing using the MaltParser trained based on information about the lemma, SUC tag and Universal PoS tag. Since the corpus format allows several types of annotation by including additional columns, scholars in Swedish studies can easily choose between them or choose to have all available annotations. When doing so, it is important to keep in mind that the same PoS tagger model should be used as that selected during the training of the parser.

5.6. Annotation Interface

One of the goals of the project is to allow researchers in the humanities and social sciences to annotate their own text and create their own corpus. For this purpose, we developed a web-based interface which allows the user to upload a text file and receive the text so that it is morpho-syntactically annotated automatically. The standard pipeline by default consists of tokenization and sentence segmentation, part-of-speech tagging with morphological features, and dependency parsing. Each module may include several algorithms and models depending on the corpus data the models were trained on. We include the most frequently used models with the highest accuracy on standard Swedish, which have been evaluated and published previously. When choosing syntactic annotation (the parser and parser model), only the PoS model that the parser was trained on may be run during the PoS tagging module to get consistent annotation.

In addition, the user may also choose to add a normalization module consisting of spelling correction and compound detection when appropriate. If so, the linguistic annotation is based on the corrected, normalized forms. The output format is the same as that described in Section 4. However, since not all fields might be relevant, the user can choose which fields (columns) are to be printed in the output file. Figure 2 shows a screenshot of the web-based interface, which can also be found at http://stp.lingfil.uu.se/swegram/. The format with fields separated by tabs allows users to import their file in Excel or another tool of their choice to carry out quantitative analysis of their choice. Furthermore, the annotation can be corrected by the user since there are annotation tools available using the CoNLL format, for example Webanno (Yimam et al., 2014), a web-based and visually supported system for distributed annotations developed for the linguistic annotation of corpora.

6. Analysis of Student Essays

Given the linguistic analysis and the structured data, the corpus can be used for a large-scale quantitative analysis of the student essays by preferred groupings based on the grade awarded, gender, and/or text type to compare the various groups with each other and investigate the development of student writings over time.

To make the analysis of student writings easier for scholars, we developed a graphical interface in Java (for platform independence) specifically for this task. The tool allows the user to read and store each text along with its metadata before performing the analysis. We also allow a combination of features so various PoS statistics are included, showing the frequency of each PoS represented in the tagset of the corpus. One feature enabling a search for sequences of PoS tags of the user's choice has also been incorporated, as well as frequency lists for each word and lemma.

For each analysis the user has the option of limiting the types of texts included. The texts can be filtered by grade awarded, gender, or whether the text is written by a student studying Swedish or Swedish as a second language. Several analyses with different filters can also be applied at the same time in order to compare different groups of students, such as showing the PoS distribution among male students and female students who have been awarded a certain grade on their text. An example of a search is given in Figure 3. There is also the possibility of adding various measures based on a combination of linguistic analyses, such as readability measures.

[5] https://github.com/robertostling/efselab

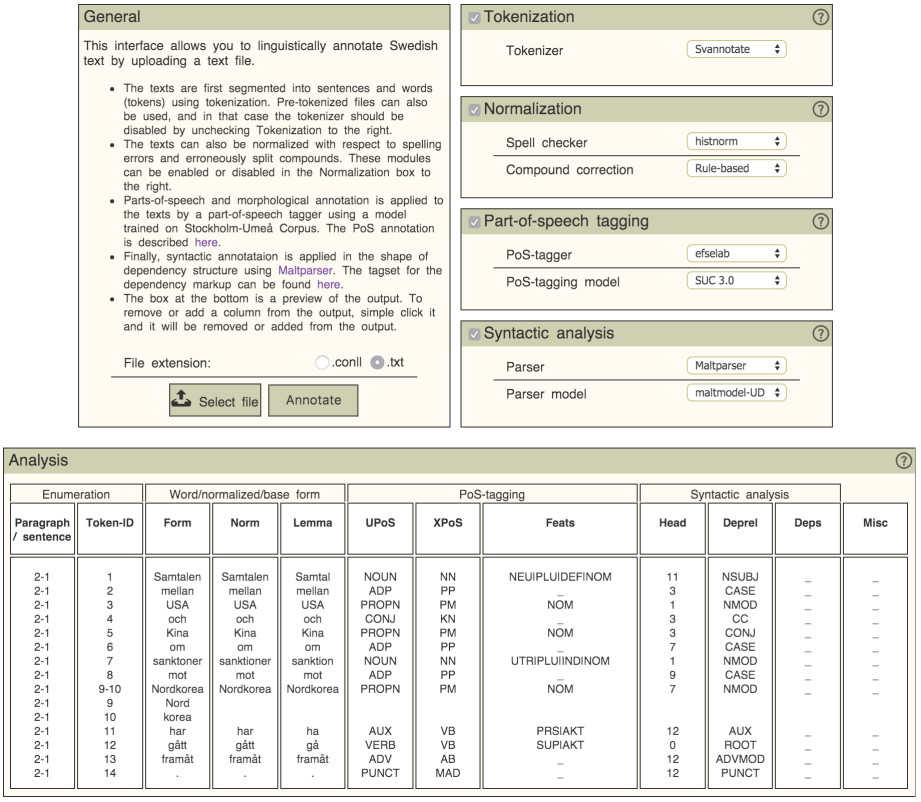

Figure 2: Screenshot of the web-based annotation interface.

Grade	**A** (Female) 9231 tokens 15 text(s)	**C** (Female) 15446 tokens 25 text(s)	**E** (Female) 4373 tokens 7 text(s)	**F** (Female) 6180 tokens 10 text(s)
Adverb (AB)	13.28%	12.20%	11.34%	10.97%
Infinitive marker (IE)	2.05%	1.58%	2.33%	1.84%
Interjection (IN)	0.08%	0.07%	0.02%	0.15%
Adjective (JJ)	7.66%	7.26%	7.48%	6.88%
Conjunction (KN)	7.44%	7.71%	8.28%	7.70%
Noun (NN)	20.39%	21.58%	20.97%	21.88%
Participle (PC)	1.06%	1.14%	0.89%	0.95%
Pronoun (PN)	18.38%	17.60%	18.04%	18.43%
Preposition (PP)	8.41%	9.17%	8.55%	9.53%
Ordinal/cardinal number (RO)	1.29%	1.54%	1.94%	1.25%
Foreign word (UO)	0.44%	0.26%	0.32%	0.34%
Verb (VB)	19.49%	19.82%	19.80%	20.05%

Figure 3: Screenshot from the output of the analysis, showing the PoS distribution for women awarded A, C, E and F as well as the total number of tokens and number of texts.

7. Conclusion

We presented a corpus of student writings (essays) written in Swedish by native speakers of Swedish or learners of Swedish as a second language from various age groups, with different genders and grades awarded. The texts are annotated at various linguistic levels, from part-of-speech and morphological features to universal dependencies. The corpus is intended to be a monitor corpus, allowing new essays to be uploaded, and automatically processed. We use automatic, mostly data-driven natural language processing tools developed for standard Swedish such as a tokenizer and sentence segmenter, PoS tagger and dependency parser.

One of the challenges in working with student texts is that spelling and grammatical errors occur frequently, which causes problems when state-of-the-art tools trained on a standard language are applied in automatic linguistic analysis. These errors could lead to a lower level of performance in a linguistic analysis. Training NLP tools on student writings would require a freely available annotated corpus of student essays made specifically for this purpose. Unfortunately, no such corpus is available for Swedish, but we believe that the corpus presented in this study may be a first step towards this goal. To avoid an increase in the error rate of the processing tools, we include a normalization step, mainly involving spelling correction as an intermediate step after tokenization and before the linguistic analysis. Grammatical errors, such as word order or agreement errors are

currently not taken into consideration, but we would like to include them in the future.

In addition to the corpus of student writings described, we also presented a web-based interface for the automatic annotation of Swedish texts. The interface enables the user to upload a file, which is then automatically fed to a pipeline of tools for tokenization and sentence segmentation, an optional spell checking, PoS tagging and morpho-syntactic analysis as well as dependency parsing of new texts. The tools are freely available and can be used by anyone who is interested in the linguistic annotation of (Swedish) text. As new, better models for standard Swedish are presented, our intention is to include them in the interface along with the old models to allow comparative studies.

Furthermore, we presented preliminary results from a quantitative study on a pilot data set of student writings to illustrate the potential of our corpus data. While the data set used is small and does not give a reliable description of the characteristics of student writings for various age groups, from different geographical areas or with different grades awarded, large quantitative studies can easily be carried out when applied to the entire corpus. However, a more sophisticated web-based GUI is urgently needed to read the processed file and output statistics based on it, to allow extraction of statistical information based on various subsamples, and to facilitate comparisons by researchers in the humanities and social sciences.

8. Acknowledgements

This project was supported by SWE-CLARIN, a Swedish consortium in Common Language Resources and Technology Infrastructure (CLARIN) financed by the Swedish Research Council for the period 2014–2018, which is aimed at creating an eResearch infrastructure that makes language resources and tools based on language resources and language technology available and ready to use for scholars of all disciplines, particularly the humanities and social sciences.

9. Bibliographical References

Garme, B. (1988). *Text och tanke: om skrivstrategier i elevuppsatser.* Liber.

Gustafson-Capková, S. and Hartmann, B., (2006). *Documentation of the Stockholm - Umeå Corpus*. Stockholm University: Department of Linguistics.

Halácsy, P., Kornai, A., and Oravecz, C. (2007). Hunpos: An open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 209–212, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hultman, T. G. and Westman, M. (1977). *Gymnasistsvenska*. LiberLäromedel, Lund.

Källgren, Gunnel et al. (2014). *Stockholm Umeå Corpus*. Stockholm University, 3.0.

Larsson, K. (1984). *Skrivförmåga. Studier av svenskt elevspråk.* Liber.

Magnusson, U. and Johansson, S. K. (2009). Quantitativa measures on student texts. In Päivi Juvonen, editor, *Språk och lärande. Rapport från ASLA: nr 22.* Stockholm: Stockholm University.

Megyesi, B. (2008). *The Open Source Tagger HunPoS for Swedish*. Uppsala University: Department of Linguistics and Philology.

Nivre, J., Hall, J., and Nilsson, J. (2006). Maltparser. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC '06, pages 2216–2219.

Nivre, J., Megyesi, B., Gustafson-Capková, S., Salomonsson, F., and Dahlqvist, B. (2008). Cultivating a Swedish treebank. In Joakim Nivre, et al., editors, *Resourceful Language Technology. A Festschrift in Honor of Anna Sågvall Hein*, pages 111–120.

Nivre, J. (2014). Universal dependencies for Swedish. *Swedish Language Technology Conference (SLTC)*.

Nivre, J. (2015). Towards a universal grammar for natural language processing. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 3–16.

Nordenfors, M. (2011). *Skriftspråksutveckling under högstadiet.* Göteborgsstudier i nordisk språkvetenskap 16.

Öhrman, L., (1998). *Felaktigt särskrivna sammansättningar.* Stockholm University, Department of Linguistics.

Östling, R. (2013). Stagger: an open-source part of speech tagger for Swedish. *Northern European Journal of Language Technology*, 3:1–18.

Östlund-Stjärnegårdh, E. (2002). *Godkänd i svenska? Bedömning och analys av gymnasieelevers texter.* Skrifter utgivna vid Institutionen för nordiska språk vid Uppsala universitet 57.

Pettersson, E., Megyesi, B., and Nivre, J. (2013). Normalisation of historical text using context-sensitive weighted levenshtein distance and compound splitting. In *Proceedings of the 19th Nordic Conference of Computational Linguistics*, NODALIDA '13.

Pettersson, E., Megyesi, B., and Nivre, J. (2014). A multilingual evaluation of three spelling normalisation methods for historical text. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH at EACL '14, pages 32–41.

Strömquist, S. (1987). *Styckevis och helt. Om styckeindelningens roll i skrivprocessen och bruket av nytt stycke i svenska elevuppsatser.* Liber.

Yimam, S., Eckart de Castilho, R., Gurevych, I., and Biemann, C. (2014). Automatic annotation suggestions and custom annotation layers in webanno. In *Proceedings of ACL-2014, demo session*, ACL '14.