Annotating Errors in Student Texts: First Experiences and Experiments

Sara Stymne, Eva Pettersson, Beáta Megyesi
Linguistics and Philology
Uppsala University
first_name.last_name@lingfil.uu.se

Anne Palmér
Scandinavian Languages
Uppsala University
anne.palmer@nordiska.uu.se

Abstract

We describe the creation of an annotation layer for word-based writing errors for a corpus of student writings. The texts are written in Swedish by students between 9 and 19 years old. Our main purpose is to identify errors regarding spelling, split compounds and merged words. In addition, we also identify simple word-based grammatical errors, including morphological errors and extra words. In this paper we describe the corpus and the annotation process, including detailed descriptions of the error types and guidelines. We find that we can perform this annotation with a substantial inter-annotator agreement, but that there are still some remaining issues with the annotation. We also report results on two pilot experiments regarding spelling correction and the consistency of downstream NLP tools, to exemplify the usefulness of the annotated corpus.

1 Introduction

The use of automatic tools for the detection and correction of writing errors is not new, and there are many tools that can accurately correct errors in standard texts in many languages, including Swedish. However, most of the existing tools are not freely available and usually do not provide any information on the error type. Automatic grammatical correction of texts written by language learners, especially second language learners is even more problematic with various types of errors.

In order to investigate language learning processes, to give students feedback, and to develop computer-assisted language learning and teaching applications (ICALL) by using NLP tools like taggers and parsers for automatic analysis of non-standard texts, it is important to be able to identify and classify various types of grammatical errors. Data collection and analysis by creating a corpus on learner language with annotation on various linguistic layers from part-of-speech (POS) to syntactic analysis is a first step. In parallel to corpus creation, tools can be developed for the automatic processing of learner data which can be used for analysis of new texts.

In this paper we present the development of a corpus on learner language of Swedish, the Uppsala Corpus of Student Writings (Megyesi et al., 2016) by creating a normalization layer identifying erroneous constructions on top of an already existing automatic linguistic annotation. In this work humans annotate word-based errors focusing on spelling, split compounds, merged words, and simple grammatical errors. The original corpus includes 2,500 student writings from different age groups and grades, written by students who study Swedish (L1) or Swedish as a second language (L2). The group of students who study Swedish as a school language consists both of native Swedish speakers, and non-native speakers who have a good command of Swedish, and those essays thus contain texts written both by L1 and L2 speakers. We describe the creation of the annotation layer for normalization for a subset of this corpus and perform two initial experiments, exemplifying how this corpus can be used.

The corpus presented is intended to be useful for researchers in computational linguistics as well as for scholars interested in student writings and assessment of Swedish as L1 and/or L2. From a computational linguistics perspective, the data will allow us to develop, train and evaluate models for error identification and correction that are particularly geared towards student writings in Swedish, possibly also adapting the models to different age groups, levels, and for students of Swedish as L1 or L2. Being able to correct errors is also im-

portant in order to achieve good performance on downstream tasks like tagging and parsing. From a writing development perspective the normalized corpus can allow analysis of writing skills development during school years in Swedish as L1 or L2. The error identification accomplished in this corpus is also interesting from an assessment and grading perspective, and can contribute to the development of advanced computer-assisted language learning and teaching applications.

2 Related Work

In research on student writing, correctness of the text is considered as one of several aspects measuring writing development and text quality. However, there is not a simple relationship between correctness and writing development. In second language writing, for example, it is well known that correctness and complexity of language are balancing factors. When focusing on correctness the student may write a less complex text, and a text with a more complex language — showing a higher level of linguistic development — may contain more errors, see e.g. Axelsson and Magnusson (2012) or Abrahamsson and Bergman (2014)).

Learner corpora allowing research studies on language learning have been available for several languages, e.g for English (Hawkins and Buttery, 2010), Norwegian (Tenfjord et al., 2004), Italian, German and Czech (Hana et al., 2004) and (Abel et al., 2014), as well as for Swedish, such as ASU (Hammarberg, 2005), CrossCheck (Lindberg and Eriksson, 2004), Swedish EALA (Saxena and Borin, 2002) and SweLL (Volodina et al., 2016). While there are hundreds of learner corpora today for various languages, only a few of them are annotated with error types along with linguistic analysis. ASK – the Norwegian Second Language Corpus (Tenfjord et al., 2004) is one important Scandinavian source including 10 different languages, annotated for errors and partly parsed. Nicholls (2003) describes the error coding and performs an analysis of the annotation of the Cambridge Learner Corpus, consisting of texts written in English by learners. She describes a scheme for inline annotations of a comprehensive set of errors. Like us, they aim to preserve both the original text and to have a corrected version. A part of this corpus has been manually annotated with POS-tags and dependency structures, and was recently released as the Treebank of Learner English (Berzak et al., 2016).

In the spell checking and grammar checking literature, e.g. Brill and Moore (2000) or Carlberger et al. (2005), corpora with annotated errors are often used for evaluation. However, little is usually written about these annotations.

3 Corpus Data

3.1 The Uppsala Corpus of Student Writings

Megyesi et al. (2016) presented the Uppsala Corpus of Student Writings (UCSW), which consists of essays written as part of Swedish national tests for schools in the subjects Swedish and Swedish as a second language. The corpus contains essays written by students in different grades, ranging from year three in primary school (at age nine) to year three in upper secondary school (at age nineteen). The tests have been collected since 1996. The texts are digitized versions either of handwritten essays, or of printed essays that have been scanned. The full corpus consists of 2,500 essays containing more than 1.5 million tokens today but the corpus is intended to be a monitor corpus, extended with new, analyzed tests.

The texts in UCSW are annotated automatically in a pipeline using SweGram (Näsman et al., 2017), an online tool for automatic analysis of Swedish texts. The tool includes tokenization, normalization to correct spelling errors and split compounds, part-of-speech tagging, and dependency parsing. First tokenization is performed to separate sentences and tokens, using the Svannotate tool (Nivre et al., 2008). Then spelling errors are corrected by using a simple unweighted Levenshtein distance, with threshold 1 on all unknown words (Pettersson et al., 2013). Split compounds are addressed by using a set of a few rules (Öhrman, 1998). Part-of-speech tagging and morphological analysis are carried out using efselab (Östling, 2016) and dependency parsing is performed using MaltParser (Nivre et al., 2006). The analysis tools achieve state-of-the-art accuracy on standard texts with the exception of the normalizer. The corrections of spelling errors and split compounds are very noisy and far from human quality, thus necessitating work on these issues.

USCW uses an extension of the CoNLL-U format, a format which is used in the universal dependency project to represent part-of-speech, morphological information and dependency relations across languages (Nivre et al., 2016). The

		Training			Test		
Level	Age	Essays (Sw)	Essays (SwSL)	Tokens	Essays (Sw)	Essays (SwSL)	Tokens
C-3	9	50	50	13,624	36	19	4,831
C-5	11	–	–	–	29	12	6,962
C-6	12	50	49	37,718	17	7	8,554
C-9	15	49	52	54,970	30	10	17,143
US-1	16	0	50	25,087	15	4	7,719
US-3	18	–	–	–	12	4	13,493
Total		149	201	131,399	139	56	58,702

Table 1: Distribution of texts by year and Sw/SwSL.

CoNLL-U format shows one token per line, with sentences separated by a blank line. For each token, it contains text and word IDs, the token, its lemma, part-of-speech tags, and dependency label plus head. For USCW, the CoNLL-U format is extended to also handle misspellings, for which an extra column is inserted, containing the correct spelling of the original tokens.

3.2 Data Used for Error Annotation

In this work we provide an added layer of human annotation on top of UCSW. In this layer we correct errors due to spelling, split compounds, merged words, and simple grammatical errors.

For this layer we sampled texts from the full UCSW corpus. We divide the data into two parts, one larger part that we intend to use as training data for NLP tools and for analysis of errors, and a smaller part intended to be used as test data. For the training data we aimed for texts that could be expected to have many errors, in particular texts from younger students, and texts written for Swedish as a second language. The main purpose for this decision was that we wanted to have a high number of errors in the data set in order to be able to train models for error correction. For the test data we aimed at a wider and more representative selection containing student texts that have been used as benchmarks in the national tests, illustrating different levels of achievement. Table 1 describes the data in the training and test sets. C refers to compulsory school, which comprises primary and lower secondary school. US refers to upper secondary school, which is not compulsory but attended by a large majority of Swedish youths.

4 Annotation

In this section we describe the error categories that were used in the manual annotation, the annotation process, and the guidelines used. We also present inter-annotator agreement for the annotators and give a summary of the identified errors.

4.1 Error Categories

The main goal of this annotation project was to find errors due to spelling, split compounds, and merged words. When we started the work we realized that it was easy to annotate simple grammatical errors at the same time. As a starting point, we decided to identify grammatical errors that only affected single words. This mainly included morphological errors and extra words.

Spelling is together with compounds the most important error type in this project. It is an error where a word is spelled incorrectly. The annotation does not consider the difference between spelling errors due to typing errors/slip of the pen, and errors due to lack of spelling competence. We include both words that are misspelled into a non-word, like *kännislor/känslor* ('feelings'), and words that are misspelled in context, but happens to form another existing word, like *ända/enda* ('end/only'). To judge if a word is correctly spelled we use SAOL, Svenska Akademiens ordlista (2006; 2015). If a spelling is accepted by SAOL, we accept it as well. This include words with alternative spellings *idag/i dag* ('today) and words with accepted informal variants *sån/sådan* ('such') and *dej/dig* ('you' Accusative). We thus allow informal spelling versions of words, as long as they are included in the SAOL dictionary, and do not enforce a particular stylistic register on the spelling norms we use. Words that are misspellings of informal spelling variants are corrected to the informal version, i.e. *non* is changed to *nån*, not to the more formal *någon* ('someone').

We also include as spelling errors cases where a word has the wrong casing, for instance for proper names: *carolina→Carolina*, or at the beginning of sentences, and errors with punctuation within words as in abbreviations: *tex→t.ex.* (*e.g.*) or hyphenated compounds: *sand-låda→sandlåda* ('sand box'). For foreign words that are part of the Swedish text, for instance movie titles or sport

terms, we correct any wrong spellings into the correct foreign spelling, if it is known to the annotators, and mark them as foreign words. An example is the English *back flipp→back flip*. For the purpose of analysis we divide the spelling errors into two groups, **casing** errors, which only concern upper/lower case of letters, and all other spelling errors.

Split compounds are cases where words that should have been written as a closed compound has instead been written as two words: *jätte bra→jättebra* ('very good'). For words that belong to a compound but that are also misspelled or have the wrong form, we correct the spelling of each part as well. In this category we also include words that are not strictly compounds, but that needs to be merged to become correct words, like *för svar→försvar* ('defense') and *kämpa de→kämpade* ('struggled') or hyphenated cases like *schim- pans→schimpans* ('chimpanzee').

Merged words are in some sense an opposite to split compounds, involving cases where two words that are supposed to be written as individual words have instead been written as one word. Examples are *tillexempel→till exempel* ('for instance') and *iår→i år* ('this year').

Simple **grammatical** errors are in this work a grouping of some different errors that concerns individual words. We view this part of the annotation as work in progress, and do not have subcategories for these errors in the annotation, we only mark them as belonging to the group of simple grammatical errors. We restrict ourselves to one-word errors to start with. The most common type of grammatical errors are morphological errors, such as agreement errors *det är viktig→det är viktigt* ('it is important') and wrong form of words *en lite by→en liten by* ('a small village'). When two words have been confused, we annotate a switch of words: *bryr som om→bryr sig om* ('cares about'). Words that have been prolonged or otherwise marked for some kind of effect are changed into their canonical version: *såååååå→så* ('so'). While it can be debated if these cases are real errors, they are problematic for automatic tools like taggers and parsers, and thus we annotate them. Finally, we annotate extra words in the text, by removing them. This could both be due to erroneous repetition: *det var igår han han klev→det var igår han klev* ('it was yesterday he stepped') or be wrong for grammatical reasons: *är en dålig på att simma→är dålig på att simma* ('is bad at swimming'). This category of errors is quite diverse, and we view this annotation as preliminary. We believe there will be the need of further sub-classification at a later stage of the annotation project that we intend to base on already existing error annotation schemes.

There are cases where an error has more than one type. In Figure 1, the last word has both a spelling error and a morphological error, and the split compound has a misspelling of one of its components. While all types of errors are corrected in these cases, for brevity and clarity of the analysis in this paper, we will mainly count each error as one type, given preference to split compounds, merged words, and simple grammatical errors in that order.

4.2 Guidelines and Problematic Cases

To aid annotation, guidelines were put together, detailing the error categories described above, and how to annotate them. The guidelines also contained numerous examples of annotations, and discussion of some borderline cases. In this section we will give some examples of problematic cases that were discussed and how we choose to annotate them, in order to give some insight into this process.

The borderline between a morphological error and a spelling error is not always completely clear. As an example we have verb forms like *hon to→hon tog* ('she took'), where the verb has a form *to*, which does not coincide with another verb form, but rather is the informal spoken pronunciation of the past verb form *tog*. For cases like this we use the strategy to annotate this as a spelling error if the student's spelling does not coincide with another verb form, and it is not clearly a misspelled erroneous verb form. A related case is the spelling of regular past verb forms, in the informal spoken from, which coincides with infinitive: *Jag svara→Jag svarade* ('I answered'). In this case, we annotate it as a simple grammatical error, since the verb form is wrong, not the spelling.

We are annotating cases where wrong words are used. However, it is often quite hard to tell which word is wrong, and what it should be exchanged with, if there are other errors or strange formulations nearby. In Example (1), it is quite clear

P-ID	S-ID	word	Auto-correct	Manual correct	Gloss	Comment
5.5	13	är	är	är		*is*
5.5	14	bläck	bläck	bläck		*ink*
5.5	15	fäj	väj	väj		*color*
5.5	16	.	.	.		
5.6	1	När	När	När		*When*
5.6	2	Bläckfisken	Bläckfisken	bläckfisken		*octopus*
5.6	3	Mar	Mar	Mar		*feels*
5.6	4	dolig	dålig	dålig		*bad*
5.5	13	är	är	är		*is*
5.5	14-15			bläckfärg		*ink color*
5.5	14	bläck				*ink*
5.5	15	fäj		färg		*color*
5.5	16	.	.	.		
5.6	1	När	När	När		*When*
5.6	2	Bläckfisken	Bläckfisken	bläckfisken		*octopus*
5.6	3	Mar	Mar	mår		*feels*
5.6	4	dolig	dålig	dåligt	x	*bad*

Figure 1: Sample of the format used for annotation: . . . *är bläck fäj. När Bläckfisken Mar dolig* . . . ('. . . is ink color. When the octopus feels bad . . . '). Top: before annotation, bottom: after annotation.

that the preposition *av* is wrong, and that it should be exchanged to *genom*. In Example (2), however, the preposition *upp* seems wrong, but it is not clear what it should be changed to, rather the whole phrase needs to be rephrased. In such examples we do not annotate anything, since that is beyond the scope of the current project.

(1) så försvarar dom sig av (genom) att
so defends them themselves of (by) to
svälja vatten
swallow water
'so they defend themselves by swallowing water'

(2) När den sener (känner) sig hotad så
When it feels itself threatened so
sveljer (sväljer) upp (?) vatten
swollows up (?) water
'When it feels threatened it swollows up [sic] water'

Another issue is morphological errors that require some kind of long-distance information to be resolved. We decided that these should be annotated as well, if they were clear from the context, even if far away. An example is shown in (3). However, when we change these types of errors it could lead to other errors, that were originally correct in the context, as shown in (4), where a correction of the co-referring pronoun from plural to singular, means that the adjective *skrämda* will have the incorrect form, whereas it was correct in the original text. In such cases we correct the adjective as well, but mark it as a grammatical error that is a consequence of the other corrections.

(3) När bläckfisken blir rädd så
When octopus.DEF becomes scared so
sprutar dom (den) bläck.
sprays them (it) ink.
'When the octopus becomes scared, it sprays ink.'

(4) Bläckfisken är blå och de (den) blir
Octopus.DEF is blue and they (it) become
ofta skrämda (skrämd)
often scared.WEAK (scared.STRONG)
'The octopus is blue and it often becomes scared'

4.3 Annotation process

The annotation was performed by four annotators, all native speakers of Swedish. Two of the annotators are computational linguists, one is a research assistant in Swedish and one is a student on the teacher training program, specializing in Swedish.

The annotation was performed in two stages. First we had a pilot stage with two phases, then we started the final annotation of the data, which is the version described in this paper. In the first pilot phase two of the annotators started work on the annotation, largely without guidelines. Specific issues were discussed between the annotators and the authors of the paper. At this stage specific guidelines were created, as described above. One of the original annotators left the project, and two new annotators were brought into the project. After a second small pilot phase, where the now three annotators discussed some issues and problematic examples, the main annotation work could

P-ID	S-ID	word	Auto-correct	Manual correct	Gloss	Comment
2.1	8	ihela	hela	hela		*in+whole*
2.1	9	kroppen	kroppen	kroppen		*body*
2.1	10	.	.	.		
2.1	8.1			i		*in*
2.1	8.2			hela		*whole*
2.1	8	ihela				*in+whole*
2.1	9	kroppen	kroppen	kroppen		*body*
2.1	10	.	.	.		

Figure 2: Sample of the format used for annotation: . . . *ihela kroppen.* ('. . . in the whole body.'). Top: before annotation, bottom: after annotation.

start with finalized guidelines. At this stage the remaining original annotator re-annotated the texts from the pilot stage, according to the new guidelines, in addition to all annotators annotating new texts from scratch. Each text is annotated by one annotator, except for the essays used for investigating inter-annotator agreement.

The annotators are given texts in a tab-separated format with one word per line, and a newline to indicate a new sentence. For each word there is a paragraph, sentence, and word number, and then the word as written by the student, and automatically corrected by the SweGram tools (Megyesi et al., 2016). The automatic annotation is also copied to a new column, where the human annotators modify it to add their annotation. In addition we insert an empty column where comments can be added, mainly used for marking the simple grammatical error category with an *x*, to tell them apart from spelling errors. The automatic corrections were given as an aid for the annotators, but they were very noisy. An example of the annotation format is shown in Figure 1, the top part before annotation, the bottom part after annotation. As can be seen, all the automatic corrections are wrong in this excerpt. Spelling errors and grammatical errors are changed in the fifth column, and grammatical errors are also marked. Split compounds are treated by inserting a new line giving the line numbers of the sub parts, and the full compound. In case of misspellings within the compound, these are also added as corrections to the individual parts, as for *fäj→färg*. A similar procedure is used for merged words, where new lines are inserted for the sub words in the merged word. An example is shown in Figure 2. The annotators used either Microsoft Excel or a text editor to do the annotation work.

Our annotation thus contains both the original text as written by the student, with potential spelling errors, split compounds etc, and the corrected version of that text, with respect to our error categories. After the human annotation, we performed automatic POS-tagging and dependency parsing of the two versions of the text, both with the original tokens, and with the corrected tokens.

	All		−correct	
	Agree	Kappa	Agree	Kappa
A1/A2	.97	.96	.72	.65
A1/A3	.97	.96	.70	.62
A2/A3	.97	.97	.72	.66

Table 2: Inter-annotator agreement and kappa for the 6-way classification between error types or correct, including and excluding the cases where both annotators judged a word as correct.

4.4 Inter-Annotator Agreement

In this section we present results on inter-annotator agreement between the three annotators that took part in the final annotation process. In order to do this analysis, a sample of 2–3 texts each from level C-3, C-5, C-6, C-9 and US-1 were chosen, with a mix of Swedish and Swedish as a second language. In total there were 11 texts with 2923 tokens. The three annotators annotated this text independently with access to the guidelines.

First we calculated agreement and kappa (Carletta, 1996) for each pair of annotators in the final phase, for the 6-way classification of each word into one of the error categories, or correct. Table 2 shows the results of this analysis. Since the majority of words are correct, the scores are very high in all cases, but even if we exclude the cases where both annotators agreed on that a word is correct, the agreement scores are reasonably high, with a kappa value over 0.6, which is considered substantial agreement (Landis and Koch, 1977). In most cases the disagreement is between an error marked by one annotator vs no error marked by another. To exemplify this, Table 3 gives the confusion matrix for annotator 1 and 2. The picture

	Co	Spe	Gr	Spl	Me	Ca
Correct (Co)	2,138		23			2
Spelling (Spe)	2	73	3			
Grammar (Gr)	15	4	65			
Split (Spl)	3			13		
Merged (Me)	1				7	
Casing (Ca)	17					23

Table 3: Confusion matrix for annotations by annotators A1 and A2, empty cells means no such confusions.

is similar for the other pairs of annotators. We see that the biggest source of confusion is where one of the annotators have considered a word as a simple grammatical error, whereas the other annotator has considered it correct. We can also see that annotator 1 has identified errors related to casing to a much larger extent than annotator 2. For the other categories the number of confusions is relatively small. For the cases where both annotators have marked a word as being either a grammatical or spelling error, the agreement of the correction is over 93% in all cases.

Overall we find the agreement satisfactory, and believe that the guidelines together with the initial discussions among the annotators were sufficient for this project.

4.5 Error Statistics

Table 4 shows the number of each error type in the training and test data. The lower part of the table shows how many spelling and grammar errors there are for components of split compounds and merged words, e.g, the split compound *jete smart/jättesmart* ('very clever') contains a spelling error of the first component. When doing this analysis we realized that our current annotations do not identify cases where a word has both a spelling and grammatical error, as for the word *dolig/dåligt* ('bad') in Figure 1.

Overall we see that spelling errors are the most common errors in both data sets. Grammar errors are nearly as common in the training set, but far less common in the test set, which could be at least parly exected, since we have more texts from young children and Swedish as a second language in the training data than in the test data. Overall we find the number of errors in both data sets sufficient for doing further research.

	Training	Test
Total	7,189	2,074
Spelling	2,826	1,205
Grammar	2,465	336
Split	548	218
Merged	192	73
Casing	1,158	242
Split+spelling	123	35
Split+grammar	29	1
Merged+spelling	46	24
Merged+grammar	9	2

Table 4: Different error types in the annotated data.

5 Pilot Experiments

In this section we will describe two pilot experiments that shows the usefulness of the human error annotation layer of UCSW. In the first experiment we show how the training data can be used for training a simple spell checker targeting student texts. In the second experiment we show how much the errors in the corpus affects automatic NLP tools, exemplified by a tagger and parser.

5.1 Spelling Correction

We can take advantage of the human annotations of student texts, in order to train tools for solving challenges like spelling correction. In this section we describe experiments on spell checking using a relatively simple approach. First we investigate how our training data impacts performance of spell checking, then we compare the performance for different student groups.

One of the most widely explored algorithms for spelling correction is to measure the *edit distance* between an unknown word and words present in a dictionary. In our spelling correction experiments, we use a simple weighted Levenshtein edit distance approach aiming to correct misspellings in the input text. The Levenshtein distance gives an indication of the similarity between two strings, by computing the minimum number of characters that need to be inserted, deleted or substituted in order to transform one string into the other string (Levenshtein, 1966). Our approach is based on the method originally presented by Pettersson et al. (2013) for the similar task of spelling normalization of historical text, and is illustrated in Figure 3. By using a weighted Levenshtein distance, we can take advantage of the training data in the human annotation layer of UCSW.

Before any normalization attempts are carried out, the program checks the length and charac-

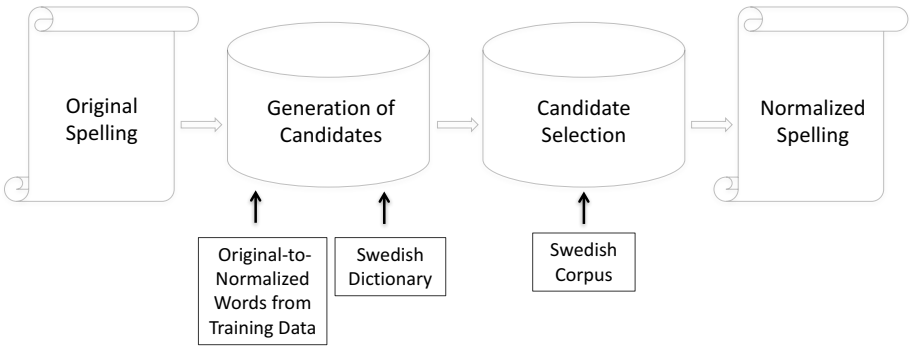

Figure 3: Flowchart for the spelling normalization procedure

teristics of the word. If the word contains only one letter, or contains digits, the word is left unchanged. Likewise, to avoid normalization of proper nouns, words with an initial uppercase letter are also left unchanged, unless they occur in sentence-initial position. One example from the test data is the string *Texten om Kissie, skriven av Malin Ekman i Expressen 10/6 2010* ('The text about Kissie, written by Malin Ekman in [the newspaper] Expressen 10/6 2010'). In this string, the proper nouns *Kissie*, *Malin*, *Ekman* and *Expressen* will be left unchanged due to their initial uppercase letters. However, the word form *Texten* ('The text') would be considered for normalization despite the uppercase first letter, since it is in sentence-initial position. The word form *i* ('in') will not be normalized due to its short length, and the date *10/6 2010* will not be normalized either, since it consists of digits.

For all word forms that do not meet these requirements, the first task is to find appropriate candidates for normalization. This is done by comparing each word form towards two lexical resources:

1. The training part of the UCSW corpus, with mappings of the students' original word forms to their manually corrected spellings.

2. The SALDO dictionary (version 2.0), a lexical resource developed for present-day written Swedish, containing approximately 1,1 million word forms (Borin et al., 2008).

If the word form is present in the SALDO dictionary, or if it occurs without having been changed in the manually normalized training part of the UCSW corpus, the word form is considered to have a correct spelling and is thus left unchanged during normalization. Else, if the word form is present in the training corpus with a normalized spelling that is different from the original spelling, this previously normalized spelling is chosen as normalization candidate. For example, the word form *henes* is not present in the dictionary. It has, however, been normalized into *hennes* (correct spelling of the pronoun 'her') in the training data. Thus, *hennes* will be chosen as normalization candidate for the word form *henes*.

If the word form is found neither in the training corpus nor in the dictionary, edit distance calculations are performed, comparing the word form to all word forms present in the dictionary. If there are dictionary entries with a Levenshtein distance of maximally one from the original word form, these entries are chosen as normalization candidates. The reason for choosing one as the maximum edit distance allowed, is that previous corpus studies have shown that misspellings usually do not differ from the intended word form with more than one edit (Kukich, 1992).

To further adapt the spelling correction process to the task of normalizing student writings, weights lower than one are included for frequently observed edits in the training data. This method has previously proven successful for example by Brill and Moore (2000) for spelling correction, and by Pettersson et al. (2013) for spelling normalization of historical text. We adopt the same approach as Pettersson et al. (2013). Thus, we split the training corpus into 90% training and 10% tuning, where the training part of the corpus is used for extracting edits to consider, by automatically comparing the historical word forms to their modern spelling, using traditional Levenshtein edit distance comparisons. The edits extracted from the

training corpus are then weighted based on their relative frequency in the tuning corpus.

One example from the corpus of student writings is the replacement of *o* by *å*, which is given the weight 0.87, meaning that it is more likely that the system will choose to replace *o* by the phonologically similar *å*, than to replace it by for example *p*. Weights for sequences of two characters on the source and/or target side are also included, mainly resulting in weights for transforming double consonants (such as *mm*) into a single occurrence of the same consonant (*m*), or the other way around.

Once the normalization candidates have been generated, a final normalization is to be chosen. This is done based on corpus statistics, in this case based on the Stockholm Umeå Corpus (SUC, version 2.0) of text representative of the Swedish language in the 1990s (Ejerhed and Källgren, 1997), containing approximately 1,2 million words. If several normalization candidates share the same minimum edit distance to the original word form, the word form with the highest frequency in the corpus is chosen. If several candidates are equally frequent in the corpus, or if none of the candidates occur in the corpus, the final normalization candidate is randomly chosen.

The above described method presupposes access to training data in the form of manually normalized student writings. If no such training data is available, spelling correction using Levenshtein calculations is still possible. In this case, the only lexical resource available during the generation of normalization candidates is a Swedish dictionary. Furthermore, traditional, unweighted Levenshtein calculations are then performed, where each edit has the cost of 1.

We present results both for the case where no training data is available (basic), and for the refined, weighted model (refined). We report results in terms of precision, recall and normalization accuracy, when running the Levenshtein-based spelling correction approach on the evaluation part of the UCSW corpus. In this evaluation setting, precision and recall are calculated for the *identification* of misspellings, that is the instances where the algorithm has correctly identified that some kind of normalization should be performed. Normalization accuracy on the other hand refers to the *correction* of misspellings, and is calculated as the percentage of correct normalizations for the true positives.

	Precision	Recall	Accuracy
basic	84.9	57.6	70.9
refined	80.9	63.5	78.2

Table 5: Spelling correction results.

5.1.1 Results on All Data

Table 5 shows the results for the spell checking. The refined method, not surprisingly, yields a higher recall, meaning that there are fewer instances of misspellings that have been left unchanged. Furthermore, normalization accuracy also increases when weights are included in the process, meaning that a larger proportion of the misspellings get an adequate correction by the refined approach. However, precision drops to some extent for the refined method. A closer look at the false positives that are unique for the refined method as compared to the basic method shows that this is almost exclusively due to real word errors in the training data. For example, the training data contains the correction of the misspelling *knakade* into the correctly spelled *knackade* ('knocked'). This means that for the refined method, having access to mappings of misspellings to their corrected forms, whenever the word form *knakade* occurs, it will be automatically changed into *knackade*. The problem is that *knakade* could also be a perfectly correct Swedish word meaning 'creaked', which results in a potential *real word error*.

Analyzing the refined correction approach further, the results table shows that about 64% of the misspellings are identified and normalized by the system. Among the false negatives, real word errors such as *varan* ('the product') vs *varann* ('each other') and *sätt* ('manner') vs *sett* ('seen') are very common. To deal with these, one would need to include context- or grammar-aware spelling correction techniques. Another common reason for false negatives to appear is that the original word form has an edit distance larger than one to the intended word form, such as *balletdansöz* vs *balettdansös* ('ballet-dancer') and *piamas* vs *pyjamas* ('pyjamas'). One way to handle these would be to experiment on different thresholds for the maximum edit distance allowed, possibly normalizing the threshold by word length.

Regarding the false positives, that is, the correctly spelled word forms that have been normal-

ized by the system even though they shouldn't have been, about a fourth of these (47 out of 181 in total) are proper nouns in sentence-initial position. Thus, more sophisticated named entity recognition would be very useful. There are also some inconsistencies in the manual normalization of the training and test corpora, which affects the number of false positives. For example, in the training part of the corpus, the word form *sej* (informal spelling of 'oneself') has been manually corrected into the more formal spelling *sig* of the same word form, which is in conflict with our guidelines. This means that system will always choose *sig* as normalization for the word form *sej*. However, in the evaluation part of the corpus, the informal spelling has been left unchanged in the manual normalization process. The same goes for the ampersand sign (*&*) and the abbreviated form *o*, which have been mapped to the word form *och* ('and') in the training part of the corpus, but have been kept unchanged in the evaluation part of the corpus. Another aspect leading to an increase in the number of false positives is the occurrence of English text within the otherwise Swedish text, which is not recognized by the system.

If these instances are ignored, about two thirds of the false positives remain as words incorrectly defined as misspellings by the system (120 instances out of 181), mainly due to a lack of coverage in the dictionary for example for compounds such as *snöhäst* ('snow horse') and *elefantben* ('elephant bones').

5.1.2 Results for Different Groups

The UCSW corpus contains texts written by different kinds of writers; younger and older students (from the age of 9 up to the age of 19), and writers studying Swedish or Swedish as a second language as school subjects. To be able to study further the kinds of errors made by the different types of writers, the training and evaluation corpora have been divided into four subcorpora:

1. Writers of all ages, studying Swedish as a school subject
2. Writers of all ages, studying Swedish as a second language
3. Younger students: from the age of 9 to the age of 12
4. Older students: from the age of 15 to the age of 19

	Prec	Recall	Acc
Swedish			
in-domain data	82.1	62.0	75.4
all data	77.9	64.2	80.7
Swedish as a second language			
in-domain data	86.2	61.9	72.0
all data	86.3	62.6	73.3
Younger students			
in-domain data	91.0	64.9	76.0
all data	87.4	65.2	76.9
Older students			
in-domain data	72.8	59.5	82.8
all data	70.0	60.2	84.6
All texts	80.9	63.5	78.2

Table 6: Spelling correction results for subparts of the evaluation corpus. Prec = Precision. Acc = Normalization Accuracy.

Table 6 shows the spelling normalization results for the different types of writers in the corpus, where experiments have been performed for training on the training corpus as a whole (referred to as 'all data' in the table) and for training on the specific subcorpus only, for example only second language training data for second language test texts (referred to as 'in-domain data' in the table).

As seen from the results, using only in-domain training data generally leads to a higher precision, due to a lower quantity of correctly spelled word forms being erroneously normalized (false positives). This is, however, at the cost of slightly lower recall and normalization accuracy, since the system then has access to less examples of correctly spelled word forms to choose from, both in the mapping of original word forms to correctly spelled word forms, and when generating normalization candidates.

It could also be noted that the system has both the highest precision and the highest recall for detecting errors in texts written by young children (age 9 to 12). Studying the misspellings in this group closer, one could see that the younger children often make errors that do not result in real word errors, and are thus recognized by the system as misspellings, such as:

- writing one consonant instead of the intended duplicated consonant, as in *fladdermös* instead of *fladdermöss* ('bat') and *överaskning* instead of *överraskning* ('surprise')
- writing duplicate consonants instead of the intended single one, as in *tännka* instead of *tänka* ('to think') and *hellikopter* instead of *helikopter* ('helicopter')

- confusing phonetically similar spellings, as in *bet**å**ng* instead of *bet**o**ng* ('concrete') and *s**c**oter* instead of *s**k**oter* ('scooter')

- writing words the way they think they sound, as in *skriskor* instead of *skri**d**skor* ('skates') and *sovenirer* instead of *so**u**venirer* ('souvenirs')

The young students also tend to use frequently occurring, common words that are often found in the dictionary when spelled correctly, resulting in relatively few instances of false positives.

The older students on the other hand (age 15 to 19 typically use less frequent and more complex word forms, that are often not found in the dictionary, such as:

- compounds, such as *regeringskritik* ('criticism against the government') and *stillös* ('lacking style')

- words that have (rather) recently entered the language, such as *chattar* ('chat groups') and *surfplatta* ('tablet device')

- slang, such as *ocoolt* ('not cool')

- abbreviated word forms, such as *o* instead of *och* ('and') and *iaf* instead of *i alla fall* ('in any case')

Interestingly though, for the word forms that have correctly been identified as misspellings, the system is better at correcting (i.e., has a higher normalization accuracy for) the texts written by older students. One reason for this is that since the older students often write less frequent and longer words, there are typically only one word in the dictionary with an edit distance of one to the original word form. For texts written by younger students on the other hand, shorter words are often used, where there are several entries to choose from as normalization candidates in the dictionary. To improve accuracy for these cases, it could be helpful to add knowledge about phonetics to the normalization algorithm (Toutanova and Moore, 2002), so that the system becomes aware that it is more likely that for example *cyckeln* should be normalized into *cykeln* ('the bike'), rather than *nyckeln* ('the key'), even if the two candidates both are within one edit distance from the original word form. Another reason that the texts written by the younger students are harder to correct is that the younger students, more often than the older students, make several mistakes for the same word form, for example when writing *jik* instead of *gick* ('walked'). Here, *j* should be replaced by *g* and *k* by *ck*. Since the generated weights for frequently observed edits are not as low as 0.5, misspellings requiring more than one edit to be corrected into the intended word form are out of the scope for the current setting.

	POS	Labels	Heads
Correct	447 (.4)	2,989 (2)	9,551 (8)
Spelling	942 (34)	994 (36)	887 (32)
Grammar	434 (16)	726 (26)	749 (27)
Split	109 (20)	247 (45)	316 (58)
Merged	108 (57)	144 (75)	139 (73)
Casing	96 (9)	138 (12)	209 (19)

Table 7: Number (percent) of confused POS-tags dependency labels and dependency heads for different error types and correct words in the training data.

The second language learners seem to make similar mistakes as the younger children, such as confusing phonetically similar spellings (such as *slengde* instead of *slängde* ('threw away')) and writing single consonants instead of duplicate ones or the other way around (such as *hottelet* instead of *hotellet* ('the hotel')). One difference is, however, that the second language learners in this corpus tend to make more mistakes related to inflection, such as writing *tågar* instead of *tåg* ('trains'), where the *-ar* ending is a more common pattern for plural inflection than the null inflection that is correct for this particular noun. This should have been annotated as a grammar error, however.

5.2 Quality of Tagging and Parsing

In this section we describe a small experiment where we compare the part-of-speech tags and dependencies automatically assigned to each word before and after the manual annotation. For this experiment we use the training corpus. Tagging was performed using efselab (Östling, 2016) and dependency parsing using MaltParser (Nivre et al., 2006). The tag sets used are the universal dependency sets both for POS and dependencies (Nivre et al., 2016). The purpose is to investigate the influence of error correction on tagging and parsing quality. Note that we do not have a gold standard for tagging and parsing, we only note how the tags change between the two conditions, not if they are correct in either case. We suspect that the tags are

POS-tag		Dependency label	
VERB-NOUN	170	nsubj-dobj	130
ADV-NOUN	152	dobj-nsubj	108
PRON-DET	90	nmod-dobj	105
ADV-ADJ	90	dobj-nmod	91
AUX-VERB	88	nsubj-nmod	72
PROPN-NOUN	85	root-advcl	70
ADJ-NOUN	81	root-nsubj	66
VERB-ADJ	81	nsubj-det	61

Table 8: The most commonly confused POS-tags and dependency labels before and after error correction.

more correct after human error annotation, however, and this is supported by a small manual inspection.

First we perform an analysis separately for each error type and correct words to see how many, and how many percent of the tokens in each category that are affected. For split compounds and merged words we compare the tag for the full word with the tag for the final word when split, which is the head word of a compound. While this is somewhat of a simplification, it can still give some idea of the influence on tagging and parsing. Table 7 shows the results. There are overall many confusions for both tools, indicating that errors indeed do cause problems for these tools. We can see that in all cases parsing is more influenced than tagging, both for predicting the correct label and for predicting the head of each word. This can in part be caused by the size of the tag sets, since there are 17 universal POS-tags and 37 universal dependency labels. While the erroneous words are affected more than the correct words, also correct words are affected by error correction. Split compounds and merged words have a very high number of confusions, which can partly be explained by our simplifying assumption of heads, but it also seems that these error types are difficult to handle for automatic tools.

Table 8 shows the most common confusions, across all error types and correct words, for POS-tags and dependency labels. The most common cases are confusions between nouns and verbs, and between subjects and objects. These are distinctions that are vital for a correct interpretation of a sentence, which again stresses the importance of good tools for error correction. There are also a high number of dependency errors involving the root of the sentence, which is also problematic. All in all the error types are quite mixed, and there is also a long tail of less common confusions.

6 Discussion and Future Work

We think the described error annotation layer is useful, but there are also some remaining issues. The spelling, split compound and merged word annotation seems to be quite sufficient, except for consistency issues with the annotation of casing. The grammatical error classification, on the other hand, would need a further sub-classification to be largely useful. In future work we also wish to handle more complicated types of grammatical errors, such as word order errors and missing words. In order to handle these errors we also need to update the format used in the corpus. It would be desirable that these annotations are consistent with previous error annotations carried out for other languages to allow cross-lingual studies.

We aim to have a single annotation scheme that covers both Swedish as a school subject and Swedish as a second language. This facilitates future comparative studies and the creation of tools for error correction. However, it is possible, especially if the scheme is extended to more complex error types in the future, that we need to have specific error types for the two variants of Swedish, since L2 language can both be expected to have more deviations from the standard norm, and have more cases with different possible interpretations of an error. Additionally we do not consider either the cause of errors, or how serious the errors are in the current annotation scheme. These are also issues that are interesting to investigate.

The analysis of the sources for issues with the spelling correction, and to some extent the inter-annotator agreement study, also pointed to some issues with the consistency of the annotation, even though the overall agreement between annotators is substantial. We thus believe that our guidelines should be extended to cover cases that were inconsistent, like the decision on the correction of casing problems. Other issues were due to annotators not following the guidelines. Yet another issue that we have noted is that we currently have no markup for combined spelling and grammar errors, which would be desirable. These issues need to be corrected in order for the annotation layer to have a high quality, which will mean we need to do more human annotation work.

This paper describes ongoing work, and we plan to annotate more texts in the future. Specifically we wish to have a more even distribution also in the training data, which would allow us to do more

comparative studies.

In this paper we described an experiment on spell checking. The approach was relatively simple, however, and we plan to use more sophisticated techniques in the future, and also to address real word spell checking. In addition we have already started work on approving the identification and correction of split compounds, and we also plan to address merged words in the future.

7 Conclusion

In this paper we have described an effort of human annotation of word-based writing errors in student texts. We described the annotation process and guidelines used in the annotation. We found that we could have a relatively high inter-annotator agreement using these guidelines. However, our analysis shows that there are still some inconsistencies in the corpora, that needs to be addressed in future work. We described a small experiment on spelling correction, to show the usefulness of the annotated corpus both for developing NLP tools like spell checkers, and for analyzing errors performed by different student groups. We also showed that errors have a large effect on POS-tagging and dependency parsing.

Acknowledgement

We would like to thank Malin Mark and Caroline Warnqvist for annotation work and Jesper Näsman for help with the SweGram tools. We also want to thank the two reviewers for their insightful comments. This work was supported by SWE-CLARIN, a Swedish consortium in Common Language Resources and Technology Infrastructure (CLARIN) financed by the Swedish Research Council 2014–2018, and SweLL financed by The Swedish Foundation of Humanities and Social Sciences 2017-2020 (IN16-0464:1).

References

Andrea Abel, Katrin Wisniewski, Lionel Nicolas, Adriane Boyd, Jirka Hana, and Detmar Meurers. 2014. A trilingual learner corpus illustrating European reference levels. *RiCOGNIZIONI. Rivista di lingue, letterature e cultura moderne*, 2(1):111–126.

Tua Abrahamsson and Pirko Bergman. 2014. *Tankarna springer före: att bedöma ett andraspråk i utveckling*. Liber, Stockholm, Sweden.

Monica Axelsson and Ulrika Magnusson. 2012. Forskning om flerspråkighet och kunskapsutveckling under skolåren. In *Flerspråkighet: en forskningsöversikt*. Vetenskapsrådet, Stockholm, Sweden.

Yevgeni Berzak, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza, and Boris Katz. 2016. Universal dependencies for learner English. In *Proceedings of the 54th Annual Meeting of the ACL*, pages 737–746, Berlin, Germany.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2008. SALDO 1.0 (Svenskt associationslexikon version 2). Språkbanken, University of Gothenburg.

Eric Brill and Robert C. Moore. 2000. An improved error model for noisy channel spelling correction. In *Proceedings of the 38th Annual Meeting of the ACL*, pages 286–293, Hong Kong.

Johan Carlberger, Rickard Domeij, Viggo Kann, and Ola Knutsson. 2005. The development and performance of a grammar checker for Swedish: A language engineering perspective. In Ola Knutsson. 2005. *Developing and Evaluating Language Tools for Writers and Learners of Swedish*. Ph.D. thesis, Royal Institute of Technology (KTH), Stockholm, Sweden.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Eva Ejerhed and Gunnel Källgren. 1997. Stockholm Umeå Corpus. Version 1.0. Produced by Department of Linguistics, Umeå University and Department of Linguistics, Stockholm University.

Björn Hammarberg. 2005. Introduktion till ASU–korpusen, en longitudinell muntlig och skriftlig textkorpus av vuxna inlärares svenska med en motsvarande del från infödda svenskar. Institutionen for lingvistik, Stockholms universitet, Sweden.

Jirka Hana, Alexandr Rosen, Svatava Škodová, and Barbora Štindlová. 2004. Error-tagged learner corpus of Czech. In *Proceedings of the Fourth Linguistic Annotation Workshop*, Uppsala, Sweden.

John A. Hawkins and Paula Buttery. 2010. Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, 1(01):1–23.

Karen Kukich. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys (CSUR)*, 24(4):377–439.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Janne Lindberg and Gunnar Eriksson. 2004. Crosscheck-korpusen – en elektronisk svensk inlärarkorpus. In *Proceedings of the ASLA Conference 2004*.

Beáta Megyesi, Jesper Näsman, and Anne Palmér. 2016. The Uppsala Corpus of Student Writings - corpus creation, annotation, and analysis. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia.

Diane Nicholls. 2003. The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 conference*, pages 572–581, Lancaster, UK.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, pages 2216–2219, Genoa, Italy.

Joakim Nivre, Beáta Megyesi, Sofia Gustafson-Capková, Filip Salomonsson, and Bengt Dahlqvist. 2008. Cultivating a Swedish treebank. In *Resourceful Language Technology: Festschrift in Honor of Anna Sågvall Hein*, pages 111–120. Acta Universitatis Upsaliensis, Uppsala, Sweden.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsafarty, and Daniel Zeman. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia.

Jesper Näsman, Beáta Megyesi, and Anne Palmér. 2017. Swegram – a web-based tool for automatic annotation and analysis of Swedish texts. In *Proceedings of the 21st Nordic Conference on Computational Linguistics (NODALIDA'17)*, Gothenburg, Sweden.

Lena Öhrman. 1998. Felaktigt särskrivna sammansättningar. Bachelor thesis, Stockholm University, Stockholm, Sweden.

Robert Östling. 2016. Shallow learning for sequence tagging. Presented at *The 6th Swedish Language Technology Conference (SLTC16)*, Umeå, Sweden.

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2013. Normalisation of historical text using context-sensitive weighted Levenshtein distance and compound splitting. In *Proceedings of the 19th Nordic Conference on Computational Linguistics (NODALIDA'13)*, Oslo, Norway.

Anju Saxena and Lars Borin. 2002. Locating and reusing sundry NLP flotsam in an e-learning application. In *Proceedings of the Workshop on Customizing knowledge in NLP applications: strategies, issues, and evaluation (LREC12)*, Las Palmas, Canary Islands, Spain.

Svenska Akademiens ordlista. 2006. *13th edition*. Svenska Akademien, Stockholm, Sweden.

Svenska Akademiens ordlista. 2015. *14th edition*. Svenska Akademien, Stockholm, Sweden.

Kari Tenfjord, Paul Meurer, and Knut Hofland. 2004. The ask-corpus - a language learner corpus of Norwegian as a second language. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal.

Kristina Toutanova and Robert Moore. 2002. Pronunciation modeling for improved spelling correction. In *Proceedings of the 40th Annual Meeting of the ACL*, pages 144–151, Philadelphia, Pennsylvania, USA.

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016. SweLL on the rise: Swedish Learner Language corpus for European Reference Level studies. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'16)*, Portorož, Slovenia.