DEPARTMENT OF PHILOSOPHY, LINGUISTICS AND THEORY OF SCIENCE

SWELL LIST

A list of productive vocabulary generated from second language learners' essays

Lorena Llozhi

Master's Thesis:	15 credits
Programme:	Master's Programme in Language Technology
Level:	Advanced level
Semester and year:	Spring, 2016
Supervisor:	Elena Volodina, Ildikó Pilán
Examiner:	Benjamin Lyngfelt
Report number:	
Keywords:	second language learning, NLP, corpus, frequency list

Abstract

Corpora for second language (L2) learning may contain a receptive vocabulary, i.e., vocabulary that is understandable by learners or productive vocabulary that L2 learners themselves are able to actively use. Corpora containing productive vocabulary could assist both students and teachers, e.g. tracking the actual learning progress, as well as language technologists who wish to analyse L2 learners' language.

While there exist productive vocabulary lists in other languages, such as the English Vocabulary Profile list, none have been made for Swedish. In this paper, we describe our project to create a Swedish vocabulary list generated from a learners' corpus, which consists of a number of second language (L2) learner essays collected into an electronic corpus. The list, named SweLL-list, contains normalised lemma and part-of-speech tag combinations and their frequency counts.

We present the work that was done to create a part of this learner corpus and the list based on it. Furthermore, we detail a normalisation algorithm, based on Levenshtein distance, used to correct L2 word level errors. We then proceed to describe our list in detail and analyse this resource through a comparison to SVALex, a vocabulary list based on L2 reading comprehension materials. Finally we examine the results of the aforementioned normalisation algorithm.

From examining the SweLL-list and comparing it to SVALex, we got some indications on the L2 students' progress. For example, we saw that while a great part of the vocabulary is taught at the intermediate levels, the students' productive vocabulary does not increase accordingly until the proficient levels.

Our analysis of the performance of Levenshtein distance for correcting L2 word level errors showed promise, especially for longer words (more than 4 characters) and where only one spelling error had been made. In order to improve the normalisation for multiple errors and shorter words, more work is needed, possibly combining the Levenshtein distance with other language technology tools.

Contents

1 Introduction

In this project, we have created a vocabulary list generated from a learners' corpus, which consists of a number of of second language (L2) learner essays collected into an electronic corpus. We have named this list SweLL-list, which stands for Swedish Learner Language (Volodina et al., 2016). Each entry in the SweLL list consists of a lemma and a part-of-speech (POS) combination and their frequency counts. The lemma (the base form of the word) and its POS have been chosen as the entries for this list, in order to make it easy to compare with other vocabulary lists. For each entry, we have included the frequencies in the corpus as a whole, both raw and normalised, as well as frequencies for each CEFR level (Council of Europe, 2001), based on the level of the essays they appear in.

The Common European Framework of Reference (CEFR) has been widely spread across Europe and provides six language development levels: A1, A2 (Basic User) B1, B2 (Independent User) C1, C2 (Proficient User) (Council of Europe, 2001:24). Among other things, the CEFR provides descriptions of the vocabulary skills that a second language (hereafter referred to as L2) learner needs to have at a corresponding level. As an example, we present here a description of the vocabulary range that a learner needs to have in order to be classified at the B2 level:

> *"Has a good range of vocabulary for matters connected to his/her field and most general topics. Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution."*
>
> Vocabulary range at B2 (Council of Europe, 2001:112)

As we can see, this descriptor at B2 level is rather vague. For instance, "a good range of vocabulary" is rather general and can have various interpretations. This vagueness has been criticised by a number of L2 language researchers, assessors and practitioners (François et al., 2016). As a result, in a number of European languages there have been various attempts to interpret the CEFR descriptors in a more concrete linguistic agenda[1] but for the Swedish language this work is lagging behind.

Due to the recently collected Swedish L2 learner essays graded for CEFR levels (Volodina et al., 2016), the chance for describing lexical repertoire that L2 Swedish learners are able to demonstrate at different levels has risen. The work described in this thesis report fills in the gap in CEFR-based lexical resources describing L2 learner productive knowledge. Questions like "Which words are productive at which level?" or "How many words does a (good) L2 learner at each level know productively?" can be answered with the help of our list. Even though we are aware of the limitations of the corpus (especially its limited size), it gives important indications as to the type of vocabulary L2 learners acquire and the type of errors they make. Also important in this context is that we can form a basis for evaluating normalization strategies that can be applied to the learner essays prior to their automatic linguistic annotation.

To get this type of list, the main prerequisite is access to electronic learner texts annotated for linguistic variables and linked to the CEFR levels. Essays at levels B2 and higher have already been digitized in other projects before the thesis work, namely Tisus and SW123 subcorpora (Volodina et al., 2016). The earlier levels were not represented, but a number of hand-written essays and their respective permits have been collected at Språkbanken. Part of our work on the SweLL-list consisted of digitization and meta-annotation of the available essays at earlier levels (144 essays in total).

Once the texts were available digitally, frequency calculations were applied to generate the first version of the list. As an experiment, we applied an algorithm based on the Levenshtein distance (Levenshtein, 1966) to correct word misspellings which are richly represented, especially at earlier

1 http://www.coe.int/t/dg4/linguistic/dnr_EN.asp?

stages of language development. Eventually, normalised L2 texts should be more reliably annotated with automated tools developed for normative texts written by native speakers. Despite the fact that Levenshtein distance (LD) as an approach to spelling error normalization has been applied to Swedish L2 texts, the conclusions can be extended to any other language with respect to LD performance on L1 versus L2 texts.

We finally compared the resulting SweLL-list with SVALex (François et al., 2016), which is based on L2 reading materials, the results of which is presented in section 8.1.

We trust that this vocabulary list will be helpful for learners, teachers, linguists, as well as for researchers within L2 acquisition and assessment interested in CEFR-specific questions. Moreover, evaluation of the Levenshtein distance within L2 context should be interesting to the NLP community working with L2 word lists in NLP learners texts.

More specifically, the SweLL-list can for instance be a self-assessment tool for the learner, as through their essay production, they can notice their vocabulary development during the course of time. The SweLL list can more easily assist the learner in comparing their receptive knowledge, i.e. the words that they are able to recognize in written texts or while listening, to their productive knowledge, i.e. the words that they actually use in the essays.

The same applies for teachers, as they can track their students' lexical development, which could help them choose the right teaching material for their future lessons. Regarding the linguists and researchers, the SweLL corpus and the SweLL-list can promote further research on L2 acquisition and more specifically on L2 learners' written production. Therefore, from a computational linguistics perspective, this could be a driving force for the advancement of the automatic L2 analysis and the improvement of L2 learning and teaching materials (Volodina et al., 2016).

The main **research questions** that we set out to answer in this project, thus, are the following:

1. Which vocabulary Swedish L2 learners can demonstrate productively in writing and how it relates to the receptive vocabulary that they acquire through reading?

2. Is Levenstein distance applicable to second language learner writing as a way of normalization? Is it reliable enough to use prior to automatic annotation?

The main **contributions** of the project are:

1. The SpIn-subcorpus – a corpus of digitized Swedish L2 essays written by learners at early stages of language development, the work that has resulted in a co-authored publication at an international conference LREC (Volodina et al., 2016).

2. The SweLL-list, a descriptive list of Swedish L2 learners' productive vocabulary, reflecting distribution of lexical items over five levels of language proficiency (A1-C1).

3. Evaluation of Levenstein distance algorithm as an approach to normalization of second language learner productive writing at single word level.

The paper is structured as follows: Section 2 introduces the reader to various vocabulary lists that include the receptive and productive type of vocabulary, after which we give an account of how the vocabulary lists are used in the NLP area. We also briefly discuss current spell checkers and their performance on L2 misspellings. Section 3 describes the workflow and methodology used in this project, while section 4 presents the SweLL corpus, the metadata and the workflow of the corpus creation and annotation. In section 5, we give an account of the procedure behind the frequency extractions and in section 6, we present the normalisation of the misspellings. Section 7 describes the created SweLL-list and section 8 contains an evaluation of the error normalisation as well as a

comparison between the two lexical resources, the SweLL-list and SVALex. Finally, section 9 concludes the report with some final remarks and outlines potential future work.

2 Background

2.1 Lexical resources for learner language

In this subsection, two types of vocabulary lists are described: lists which are derived from L2 learners' production material (essays) and lists which are created from reading comprehension materials (coursebooks, etc.). The two types of vocabulary lists have one significant difference. The first type contains the productive type of vocabulary and the other one includes the receptive type of vocabulary. A good explanation of this difference comes from Nation (2001) who writes: "Receptive carries the idea that we receive language input from others through listening or reading and try to comprehend it, productive that we produce language forms by speaking and writing to convey messages to others".

Initially, the English Vocabulary Profile (EVP) is described, which is a vocabulary list derived from essays written by L2 learners of English. The rest of the lists – FLELex, SVALex and Kelly-list are based on L1 writings, such as coursebooks, newspapers, web-texts. Before proceeding to present SVALex, we briefly describe the French vocabulary list FLElex, which served as inspiration for SVALex. Finally, we describe the Kelly list, another Swedish vocabulary list aimed at L2 learners. Of these, only EVP reflects productive type of vocabulary.

Among vocabulary lists available for Swedish, we are describing SVALex and Swedish Kelly list since they have been linked to L2 Swedish. Other lists, such as Base Vocabulary Pool (Forsbom, 2006), Swedish Academic word list (Jansson et al., 2012) and Lexin (Hult, 2012) are not relevant within L2 context, and thus are not presented here.

An important characteristic of the lists described below is that all words in the mentioned lists are assigned a CEFR level.

2.1.1 English Vocabulary Profile

An existing vocabulary list resulting from the English Vocabulary Profile (EVP) project sets state-of-the-art standards for vocabulary lists aimed at learners of a second language. This list was primarily founded by the Cambridge University Press and the Cambridge ESOL (English Profile, 2011).

The EVP list focusses on the words that students already know the meaning of, rather than the words that students need to know (Capel, 2010). Various corpora were used to create the EVP, including the Cambridge Learner Corpus which contains a large number of exam scripts written by L2 learners. The EVP is also based on the Cambridge English corpus, which is a multi-billion word corpus of written and spoken English. Additional types of data were also used for the creation of the EVP such as "examination vocabulary lists, classroom materials and a wide range of course books"[2].

The EVP is browsable and freely available for everyone[3]. The user can choose between a basic or an advanced search option. The advanced option returns a detailed outline of the words' senses, contrary to the basic search which returns a limited overview of the words' senses (Capel, 2010). It is worth mentioning a few things about how the insertion of the words' senses was conducted. Initially, the Cambridge International Corpus was used for this project and the different senses of the words were

2 http://languageresearch.cambridge.org/wordlists/compiling-the-evp

3 http://vocabulary.englishprofile.org

selected based on their relative frequency among the first 5,000-6,000 words (Capel, 2010). Thereafter, “lexicographers manually counted concordance lines for these words and, according to the number of occurrences of a given sense, assigned one of the relative frequency to it: E, I and A (Essential, Improver and Advanced)” (Capel, 2010). The user can filter for words or phrases by selecting a number of features that they would like to derive, for instance, the word level according to the CEFR, an audio and written pronunciation, grammar and usage information and one or more native speaker dictionary examples. Another interesting characteristic of the EVP is that it provides “authentic examples of learner writing from the Cambridge Learner Corpus”[4]. Those examples are provided with information about the students' native language, the CEFR level of the exam and the language exam that the student took.

2.1.2 FLELex

An inspiration for SVALex was the FLELex project, a French vocabulary list designed for students who learn French (François et al., 2014). The FLELex lexicon is linked to the 6 CEFR levels and the corpus from which it was constructed included 28 textbooks. The material was selected based on two principles; firstly, it had to have been published from 2001 and onward and secondly, it had to be designed for general purpose learning.

The final corpus consists of 777,000 running words which gave rise to a FLELex list with 14,236 alternatively 17,871 unique lemmas and part-of-speech tag combinations with their respective frequency counts, the counts being different depending upon the tagger applied to the corpus for part-of-speech tagging (François et al., 2014).

2.1.3 SVALex

SVALex is a vocabulary list created from the COCTAILL corpus (Volodina et al., 2014), and reuses methodology suggested in FLELex. COCTAILL is derived from coursebooks which are used at Swedish language lessons as L2 and are linked to the CEFR levels. The corpus was linguistically annotated by the Korp pipeline (Borin et al., 2012).

More specifically, the SVALex list includes 15,681 word types and features for each word type: the lemma, part of speech (POS) and frequencies across 5 CEFR levels (François et al., 2016). Additionally, the list contains multi-word expressions which were detected in the text by using the Korp pipeline.

While calculating the frequencies of the word entries, a dispersion to raw frequencies was used in order to eliminate instances of words for which the frequency is high in only certain texts while not in the overall corpus (François et al., 2016). As a result, the frequencies of the words become more accurate and more representative of the corpus as a whole.

2.1.4 Swedish Kelly list

The Swedish Kelly list is part of a bigger project initiated by the European Union's Lifelong Learning Programme (Volodina & Johansson Kokkinakis, 2012). The aim of the Kelly project was to construct frequency word lists for each of the following languages: Arabic, Chinese, English, Greek, Italian,

4 http://vocabulary.englishprofile.org/staticfiles/about.html

Norwegian, Polish, Russian and Swedish (Charalabopoulou et al., 2012). An important aspect was that each resulting list was translated into each of the other eight languages.

The SweWac corpus was used for the construction of the Swedish Kelly list. SweWac is a web-based corpus that includes 114 million words (Volodina & Johansson Kokkinakis, 2012a). Its core constitutes L1 texts and it was annotated with tools developed by Kokkinakis and Johansson Kokkinakis (Charalabopoulou et al., 2012).

A central aim of the Swedish Kelly list was that it should reflect a contemporary type of language that is not restricted either by genre (Volodina & Johansson Kokkinakis, 2012a) or topic (Charalabopoulou et al., 2012). Additionally, it should include both words that appear often in the corpus as well as words that are deemed useful to learn by students at various levels. This means that it also should be representative of the vocabulary introduced at the different CEFR levels.

The Swedish Kelly list includes 8,425 word entries, each linked to a CEFR level based on its frequency (Volodina & Johansson Kokkinakis, 2012b). Also, a frequency measure was calculated for the equivalent POS of each word entry.

The frequency analysis of the list was conducted by using the raw frequency (RF), the relative frequency (word per million or WPM) and the average reduced frequency (ARF), in which counts for multiple occurrences that are close to each other in the text are lowered (Volodina & Johansson Kokkinakis, 2012a).

2.2 Lexical resources in an NLP context

2.2.1 Word lists in NLP

Word lists are frequently used in Natural Language Processing (NLP) applications. With regards to the Swedish language, Pilán et al (2014) used them for analysing sentence readability, while Heimann Mühlenbock (2013) did similar analyses for text readability. François & Fairon (2012) used word lists in their study exploring a new readability formula for French as L2. There are also other uses of word lists in the language technology sector, such as in machine translation, speech recognition, readability analysis, information extraction, etc.

2.2.2 Spelling error correction

The vast majority of automatic spell checkers are created for the native speakers of a language (Antonsen, 2012), and assume that the errors made by the user are “accidental typographical mistakes” (Rimrott & Heift, 2005). As a consequence, they perform well when dealing with “errors of addition, omission, substitution, and transposition”. However, non-native speakers also make errors which differ more greatly from the intended word, due to the speaker's incomplete language knowledge. Since automatic spell checkers are trained to correct native speakers' misspellings, their performance with regards to mistakes made by non-native users is not considered successful (Rimrott & Heift, 2005).

Recent studies have addressed L2 spelling errors and proposed new methods in order to improve the performance of existing spell checkers. For students learning North Sami as L2, a finite state transducer (FST) was employed with the aim of “improving the feedback on L2 misspellings”, taking the context of the word into consideration (Antonsen, 2012). Other studies have focussed on the grammatical errors made by L2 learners. These are, however, outside the scope of this paper, and we refer the interested reader to e.g. Tou Ng et al. (2014) and De Felice & Pulman (2008).

In our study, we evaluate the performance of an algorithm based on the Levenshtein distance as applied to word-level errors in the Swell corpus (Volodina et al., 2016). The Levenshtein distance measures the difference between two strings and has been used for error correction in various linguistic contexts such as historical linguistics (Hauser & Schulz, 2007) and information retrieval (Cucerzan & Brill, 2004). However, we are not aware of the studies where Levensten distance is applied to L2 writing. The main principle of Levenstein distance algorithm can be summarized as follows: An exact match will give a score of 0, while any character substitution, addition or subtraction adds one to the score. As such, the lower the score, the closer the match (Levenshtein, 1966).

3 Workflow and methodology

Our work behind the creation of the SweLL list, can be divided into four phases. In figure 1, we introduce the reader to the various steps that encompass the creation of the SweLL list:

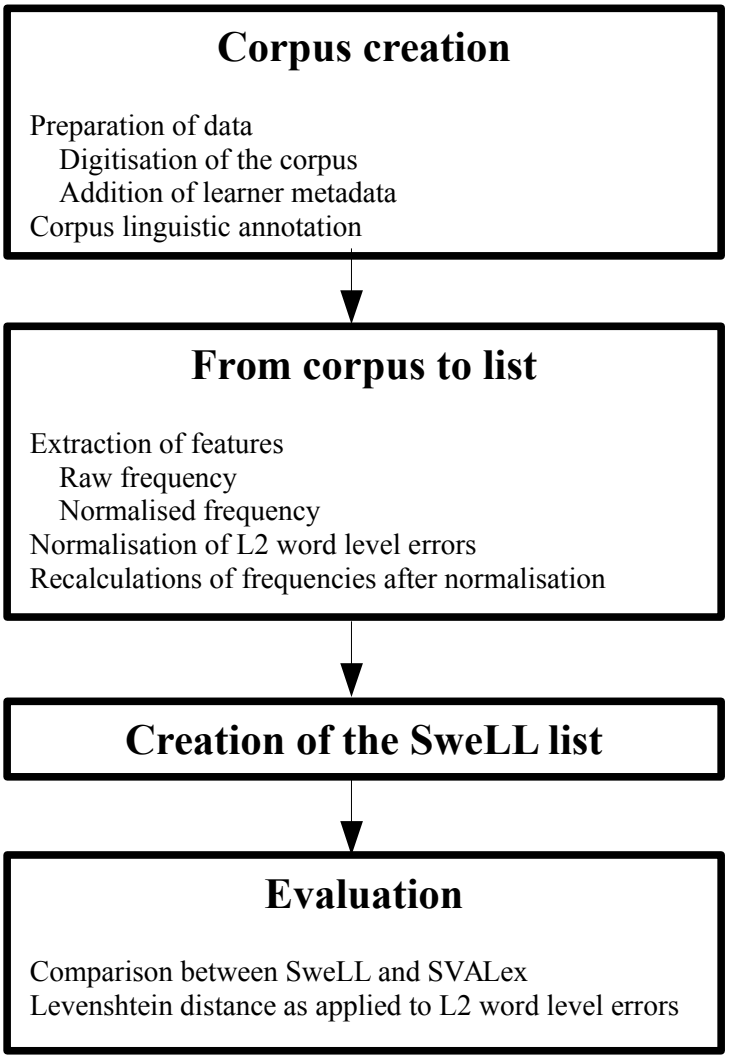

Figure 1: The workflow of the SweLL-list creation

4 The source data

4.1 The SweLL corpus

The SweLL corpus consists of 339 essays divided into three different subcorpora: SpIn, SW1203 and Tisus. In table 1, we present the number of essays per subcorpus as well as the number of essays per CEFR level (Volodina et al., 2016). As we can see, the current SweLL corpus contains essays from five different CEFR levels, namely A1-C1.

Subcorpus	**A1**	**A2**	**B1**	**B2**	**C1**	**Unknown**	**Total**
SpIn	-	-	-	27	78	-	105
SW1203	-	-	33	45	11	1	90
Tisus	16	83	42	2	-	1	144
Total	16	83	75	74	89	2	339

Table 1: SweLL subcorpora (from Volodina et al., 2016)

4.1.1 SpIn subcorpus

The SpIn corpus (Centrum för **Sp**råk**In**troduktion) consists of 144 essays collected from the Center of Language Introduction (Volodina et al., 2016). This language school accepts students between ages 16 to 20. The students are either refugees or immigrants and usually start learning the Swedish language at a basic level. The purpose of this language programme is to give the efficient knowledge to the learners in order for them to continue to the "next transitional training stage" and eventually continue with their studies at the Swedish upper secondary schools. The essays collected for the SpIn corpus are gathered from a language test that the students need to take every 7 weeks. After the test the students continue studying at the level equivalent to the achieved score at the exam. All of the essays are assigned a CEFR level. Furthermore, several students have written more than one essay during the course, which makes it possible to observe the students' language progress through the course of time.

4.1.2 Tisus subcorpus

The Tisus subcorpus contains 105 essays collected from the TISUS (Test in Swedish for University Studies) exam. Students who intend to pursue their academic studies at a Swedish university, where the studies are conducted in Swedish, need to demonstrate their Swedish skills by passing this exam. The test includes three parts; reading, speaking and writing. The final result is either a Pass (Godkänd) or Fail (Underkänd). The essays used for the Tisus subcorpus are all discussing the same subject ("Stress") and share a common genre of argumentative writing (Volodina et al., 2016).

4.1.3 SW1203 subcorpus

SW1203 comes from the course “Swedish as a foreign language – Qualifying course in Swedish”, which is provided by the University of Gothenburg. Students who intend to continue their studies at a university level, particularly studies which are held in Swedish, can take this course as part of a language training program (Volodina et al., 2016).

The SW1203 subcorpus contains 90 essays, written in three different parts of the course, including an entrance exam, an evaluation test taken in the middle of the term and a final test at the end of the course. Therefore, as in the SpIn subcorpus, we can encounter in SW1203 several essays written by the same students. This gives us the opportunity to look deeper into the students' language progress.

4.2 Digitization details and issues

As part of our project, we digitised the 144 handwritten essays in SpIn, so that they could be processed by the computer. Each essay had one of the two types of research permission, as given by the students. Either we were given permission only for restricted use, which means that only researchers included in the project can use the essays, or permission for unrestricted use, provided that the student's anonymity was preserved. During the digitization, any information that would reveal a student's identity was replaced by NN, and references to places would be replaced by e.g. N-gata (Eng: N-street).

An important feature of all of our corpora is that they contain a number of misspelled words, which is a natural feature since they were collected from L2 students' essays. None of these misspellings were corrected and the original form of the essay was preserved in the digitised version. However, we applied a positive assumption to potentially erroneous segments/words that were not clearly distinguishable and presumed that the learner had written a correct form of the word (Volodina et al., 2016). In certain cases, words or characters were not comprehensible at all. These, we replaced with one @-token for each illegible character .

4.3 Learner variables and linguistic annotation

The digitised essays included a number of metadata regarding students' profiles and their submitted essays. In order to store these metadata, we used the essay editor in Lärka, which is “a learning platform designed for learning Swedish” (Volodina & Lindström Tiedemann, 2014). A screenshot of Lärka editor can be seen in Figure 2.

When using the Lärka essay editor “an annotator is steered through prompts to fill in or values to select from” (Volodina et al., 2016). When submitting an essay, the essay editor automatically suggests an essay ID and a student ID. In cases where a student has written more than one essay then the essay ID would refer to the same student ID.

The metadata which were saved for each essay include:

- The information related to the students' profile. This included the students' gender, age, mother tongue(s), residence time in Sweden and their educational background.
- Additional information on the submitted essays. These would be the CEFR level, the setting (exam, classroom or home), usage of any additional material while writing the essay (e.g. dictionaries), the semester (autumn, spring), the date of writing and the essay's topic.
- The subcorpus in which the essay can be found.

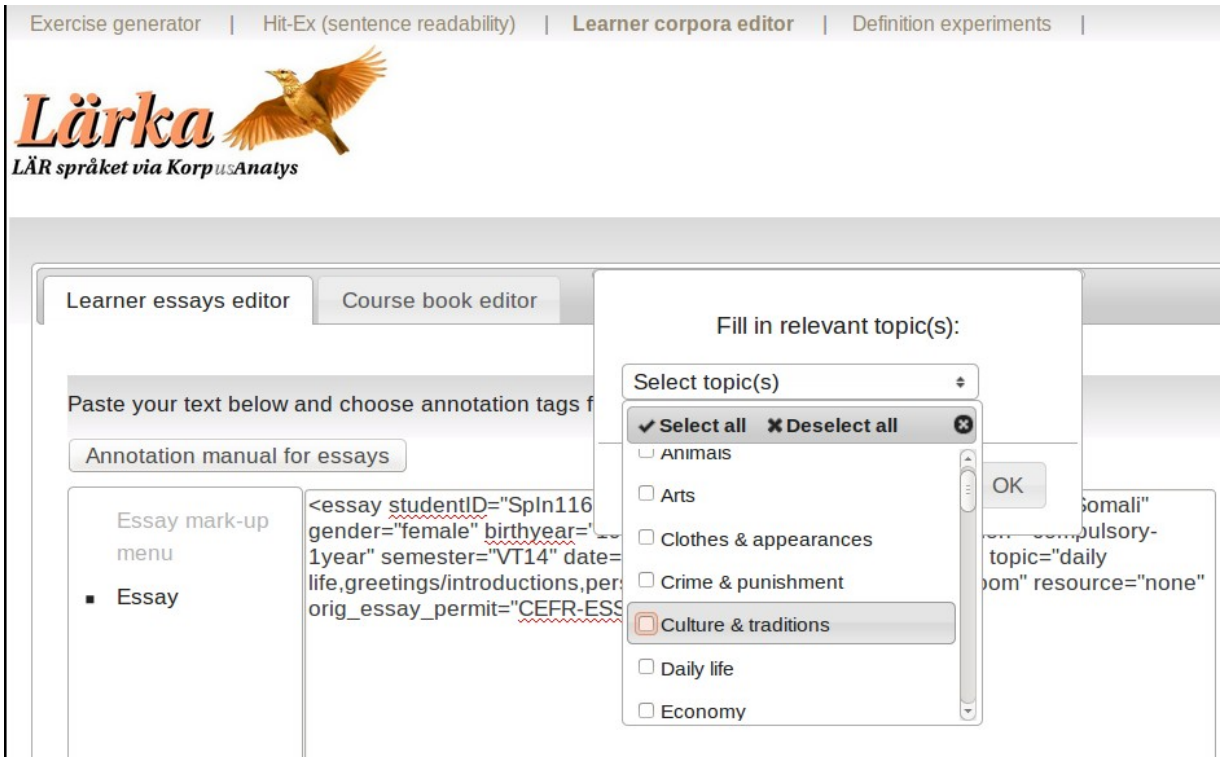

Figure 2: The Lärka essay editor

After submitting the aforementioned information, an XML-tag is generated (Volodina et al., 2016) which has the following structure (Figure 3):

```
<essay age=" " birthyear=" " cefr=" " date=" " education=" " essay_id=" " gender="
" l1=" " permit=" " residence=" " resource=" " semester=" " setting=" "
student_id=" " subcorpus=" " topic=" ">
```

Figure 3: An example of an XML-tag with essay-specific metadata

The submitted essays were then tokenised and annotated by the Korp pipeline, which is used for "importing corpora, annotate them, and then exporting the annotated corpora into different formats" (Borin et al., 2012). The annotation included the lemmatization, part-of-speech tagging and syntactic information. After an essay had been processed by the Korp pipeline, we ended up with the following XML structure (Figure 4):

```
<essay>
  <sentence>
    <w>
```

Figure 4: SweLL corpus XML-structure

Each tag has a number of attributes which were inserted either when using the Lärka tool (<essay>) or the Korp pipeline (<sentence>, <w>). Above, we have mentioned the attributes that the <essay> tag

includes after using Lärka tool. During the annotation by Korp, the following metadata are added to the <sentence> and to the <w>, i.e. word, XML-level:

The <sentence> receives an id as an attribute (Figure 5):

```
<sentence id="2bec5f17-2bb174f5">
```

Figure 5: An example of an XML-tag with sentence specific metadata

whereas the <w> tag will get the following attributes (Figure 6):

```
<w pos="VB" msd="VB.INF.AKT" lemma="|svära|" lex="|svära..vb.1|" saldo="|svära..1|
svära..2|" prefix="|" suffix="|" ref="11" dephead="10" deprel="+F">svära</w>
```

Figure 6: An example of an XML-tag with word specific metadata

These attributes can be divided into three subcategories: lexical, compound and dependency attributes. The attributes *pos, msd, lemma, lex* and *saldo* refer to the lexical analysis, whereas the *suffix* and *prefix* attributes refer to the compound analysis of the word entry. Lastly, the *ref, dephead* and *deprel* refer to the dependency analysis of the word.

During the processing of the word entry by the Korp pipeline, the attributes are either generated from the Korp pipeline or taken from SALDO, which is a "semantic and morphological lexical resource" (Borin et al., 2013) currently containing metadata for 137,130 word entries[5].

The *lemma, lex, saldo, prefix* and *suffix* attributes are the ones generated from SALDO. More specifically, the *lemma* contains the basic word form of a token while the *lex* refers to a so-called lemgram - a combination of lemma and POS tag which also identifies its inflectional paradigm. The *saldo* attribute contains the possible senses of the lemma, while the *suffix* and *prefix* tags include the initial and the final part of a compound respectively [6].

Meanwhile, the *pos, msd, ref, dephead* and *deprel* attributes are created by the Korp pipeline. The *pos* attribute is the part-of-speech while *msd* includes the morphosyntactic features of the word entry. Furthermore, *dephead* indicates the dependency that the word has to the head of the sentence, *deprel* contains the word's syntactic role (e.g subject as SS) and *ref* indicates the word's position in the sentence. An overview of the word level metadata is presented in Table 2.

Korp pipeline attributes		
Lexical analysis	**POS**	Part of speech
	msd	Morphosyntactic features
	lemma	The lemma or base form of the word
	lex	Lemgram, containing grammatical data

5 https://spraakbanken.gu.se/eng/research/saldo/statistics

6 https://spraakbanken.gu.se/eng/research/infrastructure/korp/annotations-in-korp

Korp pipeline attributes		
	saldo	The word sense
Compound analysis	**suffix**	The word's suffix
	prefix	The word's prefix
Dependency analysis	**ref**	The word's position in the sentence
	dephead	The dependency of the word to the sentence
	deprel	The word's syntactic role in the sentence

Table 2: An overview of word level metadata added in the Korp pipeline

After an essay has been analysed by the Korp pipeline and its metadata are stored by Lärka, the aforementioned attributes are populated and the essay gets a resulting XML format as shown in Figure 7:

```
<essay subcorpus="SpIn" student_id="SpIn77" essay_id="SpIn77_5" cefr="B1" l1="Mandarin Chinese, English" age="17" education="upper-secondary-3-4years" permit="public" gender="female" topic="arts" birthyear="1996" residence="7" semester="VT13" date="04-2013" setting="exam" resource="none">

       <sentence id="27ee71e-275e415">

              <w pos="NN" msd="NN.UTR.SIN.DEF.NOM" lemma="|film|" lex="|
              film..nn.1|" saldo="|film..1|film..2|film..3|" prefix="|fil..nn.2|
              fil..nn.3|fila..vb.1|fil..nn.1|" suffix="|men..nn.2|men..nn.1|
              m..nn.1|" ref="1" dephead="2" deprel="SS">Filmen</w>

              <w pos="VB" msd="VB.PRS.AKT" lemma="|handla|" lex="|handla..vb.2|
              handla..vb.1|" saldo="|handla..4|handla..1|handla..2|handla..3|"
              prefix="|" suffix="|" ref="2" dephead="" deprel="ROOT">handlar</w>

              <w pos="PP" msd="PP" lemma="|om|" lex="|om..pp.1|" saldo="|om..1|
              om..5|" prefix="|" suffix="|" ref="3" dephead="2" deprel="OA">om</w>

              <w pos="DT" msd="DT.UTR.SIN.IND" lemma="|en|" lex="|en..al.1|"
              saldo="|den..1|en..2|" prefix="|" suffix="|" ref="4" dephead="5" de-
              prel="DT">en</w>

              <w pos="NN" msd="NN.UTR.SIN.IND.NOM" lemma="|kille|" lex="|
              kille..nn.2|kille..nn.1|" saldo="|kille..2|kille..1|kille..3|" pre-
              fix="|" suffix="|" ref="5" dephead="3" deprel="PA">kille</w>

              <w pos="HP" msd="HP.-.-.-" lemma="|" lex="|" saldo="|" prefix="|"
```

```
            suffix="|" ref="6" dephead="7" deprel="SS">som</w>

            <w pos="VB" msd="VB.PRS.AKT" lemma="|heta|" lex="|heta..vb.1|"
            saldo="|heta..1|" prefix="|" suffix="|" ref="7" dephead="5"
            deprel="ET">heter</w>

            <w pos="PM" msd="PM.NOM" lemma="|Billy|" lex="|Billy..pm.1|"
            saldo="|Billy..1|" prefix="|" suffix="|" ref="8" dephead="7"
            deprel="OO">Billy</w>

            <w pos="MAD" msd="MAD" lemma="|" lex="|" saldo="|" prefix="|" suf-
            fix="|" ref="9" dephead="2" deprel="IP">.</w>

      </sentence>

</essay>
```

Figure 7: Excerpt of the XML after an essay is processed by the Korp pipeline

5 Extraction of frequencies

After the essays were processed by the Korp pipeline, the output was an annotated XML file. The next step was to extract the relevant attributes from the subcorpora and collect the frequencies. By collecting these pieces of information we would be able to create our final word list.

The attributes that we extracted for the production of our final list included the lemma, POS and MSD. Furthermore, we collected some additional attributes from the annotated essays which could also be potentially useful for tracking the students' progress. Those attributes were the CEFR level of each essay and the subcorpus, as well as the essay ID and the student ID. The processing of the data and collection of relevant information associated with each entry, was performed automatically using python programming language.

Note that multi-word expressions and the following tokens were ignored:

- tokens that were tagged as proper names and punctuation
- tokens that consisted of or contained digits
- tokens containing the percentage symbol

Our next step was to collect the frequencies of the lemma and part-of-speech combination (lemma-pos). We collected the frequencies of the lemma-pos occurrences for the five CEFR levels and for each subcorpus. We used two types of frequencies: the raw and a normalised frequency. The raw frequency is the actual number of occurrences of each lemma. For the normalised frequency we used the WPM (word per million), in which the number of occurrences is divided by the total number of tokens and multiplied by one million. The purpose of including WPM is to make SweLL-list frequencies comparable with other corpora (François et al., 2016).

At this point, the first version of our list was created. We present below a simplified version of its form in Table 3. To conserve space, we have included only A1 level and exemplify with only one essay-ID.

Lemma	allt
POS	AB
raw_freq	90
normalised_freq	773.786
A1	1
Tisus_A1	1
SweLL_A1	0
SpIn_A1	0
EssayIDs	[TISUS58_58:7]

StudentIDs	[Tisus58:7]
Tokens	[allt]
MSD	[AB]

Table 3: An overview of our final table with data from SweLL

After collecting the frequencies, we noticed that there were 4,308 unique tokens which were not assigned a lemma during the linguistic annotation. We found that there were generally three reasons for why a word was not lemmatised by the Korp pipeline: (1) either they were misspelled words, (2) they were not present in SALDO, (3) or the Korp pipeline did not manage to fetch the lemma from SALDO. Regarding the latter case, the Korp pipeline in some cases cannot recognise the lemma from SALDO as it uses a different tagset. For instance, the pos tag for the word lagom (sufficient) was marked as JJ (adjective) by the Korp pipeline and av.1 (adjective) by SALDO.

In Figure 8, the distribution of non-lemmatised tokens after annotation through the Korp pipeline is presented in the respective percentages per level:

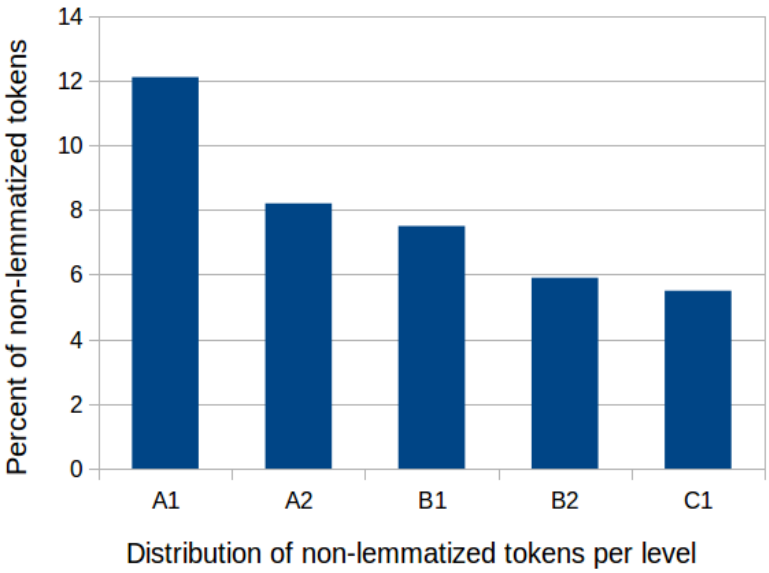

Figure 8: Distribution of non-lemmatised tokens in percentages per level (from Volodina et al., 2016)

As we can see, the rate of non-lemmatised tokens per CEFR level is quite high. The unlemmatised tokens reach the highest percentage at the A1 level (12%). This is gradually falling as we approach the intermediate and proficiency levels but stays in an overall high rate. We can assume that among the reasons why the tokens failed to be lemmatised by the Korp annotation system, the word-level errors were the main reason. From a second language acquisition point of view, the mentioned rates are reasonable since the L2 learner is more likely to make mistakes at the beginner level when they are not that well acquainted with the target language.

In order to determine the possibility of misspellings for those words where Korp pipeline could not find a lemma, we decided to check if these particular tokens were present in the Saldo morphology XML file (Saldom), which is a part of the SALDO resource. Saldom is a full-form lexicon i.e. it contains lists of all inflected forms for each baseform and has 128,036 word entries in total[7]. Thus, we wrote a script that reads the original Saldom XML into a python dictionary, which we then used to look up word forms. Our output from this script is a programming object where the word form (token) is the key, and the value is a list of paired values, so-called tuples (lemma, lemgram, pos, msd). We show an example of an entry for the token "får" ("receives, verb, present tense, singular" or "sheep, noun, base form") in Figure 9:

```
får:   [(lemma: få, lemgram: få..vb.1, pos: vb, msd: pres ind aktiv)
        (lemma: får, lemgram: får..nn.1, pos: nn, msd: sg indef nom)
        (lemma: får, lemgram: får..nn.1, pos: nn, msd: pl indef nom)]
```

Figure 9: Example entry of a word form in our dictionary

Then, we ran all the tokens with unknown lemmas through a pre-normalisation function where they were matched against morphologically inflected forms in Saldom. In the case that we found a match to a morphologically inflected form in Saldom, we chose that entry's lemma. Whenever there were more lemmas to choose from, we picked the first for reasons of simplicity.

After retrieving possible Saldom matches for the unlemmatised tokens, we found there were still 4,566 tokens (3,053 unique ones) which were not present in SALDO. We analysed a selection of these words (about 1000 tokens) and split these in five categories:

- misspelled tokens
- compound words
- words that had incorrectly been split with a hyphen, e.g. "för-söka". From the examples we examined, we assumed that it is either a student's mistake or an error during the digitization process. Although it could have been easily fixed code-wise, we decided not to interfere as we could have excluded expressions which are supposed to be written with a hyphen.
- foreign words
- acronyms

In Table 4, we show some examples of these five categories, including the correct spelling and English translation where applicable. Misspelled words are (here and elsewhere in this paper) marked with an asterisk (*).

Tokens not present in SALDO	**Examples**	**Correct spelling**	**English translation**
Misspelled words	fotbol*	fotboll	football
	rappot*	rapport	report

7 https://spraakbanken.gu.se/resource/saldom#tabs=information

compound words	projektanställning		project employment
	onödan		in vain
	arbetstempo		work pace
words split with a hyphen	väster-ländska	västerländska	western (adj)
	för-söka	försöka	attempt
	an-ställa	anställa	employ, recruit
foreign words	opportunity		
	coffee		
	success		
acronyms	p.s		post scriptum
	pgv	pga (på grund av)	due to

Table 4: Examples of word entries that we were unable to find a match for in Saldom

At this point, we proceeded with an experiment to normalise the remaining 4,566 unlemmatised tokens.

6 Normalization of word-level errors

As we expected, the deviating learners' language affected the quality of the lemmatisation and the annotation of the corpus. In order to overcome this deficiency, we decided to take a step further and add a normalisation method, based on the Levenshtein distance (LD), for word-level errors. Consequently, this project has become a test case for LD on L2 material and we later include in our evaluation the performance of this approach on errors that are specific of non-native writers.

As we previously mentioned, LD is a measurement for the distance between two strings. In our case, this would be the difference between the (possibly) misspelled word and the (probable) target word. As such, in the cases where the word form is not present in SALDO, we chose the lemma in SALDO to which the word form in our source had the shortest LD.

In order to improve the overall speed of our method and eliminate the noise (such as wrong lemma mappings) we added a prerequisite that both the word form and the potential mapped lemma should begin with the same letter. This prerequisite was based on the assumption that a misspelled word is likely to start with the correct letter of the corresponding lemma (Rimrott & Heift, 2005).

The Levenshtein distance measurer we used in this project, was already implemented and available in the Natural Language ToolKit (NLTK) Python package (Bird, 2006)[8]. Our algorithm needed to make two loops through the Saldom for each token we wanted to normalise. First, it went through all lemmas in Saldom in order to determine what the shortest possible LD to the token is. Then, when we have the shortest possible distance, a second loop is performed, and the first lemma with that distance is chosen. As we will discuss in section 8, this procedure worked well for certain tokens, but less so for others.

After applying our normalisation algorithm to the unlemmatised entries, we re-calculated the frequencies. Thus, the final SweLL-list was created, as described in section 7.

8 http://www.nltk.org/

7 Description of the SweLL-list

The final SweLL-list contains 6,805 unique items that students across A1-C1 actively use. This vocabulary corresponds to the learners' productive knowledge. The distribution of the vocabulary that the learners use across the CEFR levels is presented in Table 5. Note that *lemma* here means the lemma-pos combination.

CEFR	**# of unique lemmas**	**# of new lemmas**	**# of lemmas per level**
A1	447	447	1,646
A2	1,629	1,326	15,376
B1	2,776	1,944	26,408
B2	2,670	1,222	29,011
C1	3,970	1,866	53,687

Table 5: Vocabulary distribution in the SweLL-list per level

The *unique lemmas* column shows the number of different lemma-pos combinations that students at various levels have used. For example, at level B1 we can see that the students have used 2,776 unique lemmas in their essays. This does not mean that all the unique lemmas at B1 level were only used at this level, as it also includes lemmas overlapping with lower levels.

In contrast, the column *new items* reveals the number of lemma-pos combinations that have not been used until the respective CEFR levels. As such, these lemmas can be linked to the lexical progression of the students. We can observe that it reaches its highest peak at the B1 level (intermediate) with a new rise at the C1 level (proficiency). In other words, it is at the transition levels, from basic to intermediate and from the intermediate level to proficiency, that the highest vocabulary acquisition can be observed.

Looking closer at the actual lemmas in our list, we present the top 10 words for each level in table 6. In those cases where the same lemma appears more than once, we have included the POS tag to distinguish between them:

A1	**A2**	**B1**	**B2**	**C1**
jag_PN	jag_PN	vara	vara	en
och	och	och	en	vara
jag_PS	vara	jag_PN	och	och
skola	jag_PS	en	den	att_IE
i	på	den	att_IE	den

gå	i	i	man	att_SN
vara	till	att_IE	att_SN	i
på	en	vi	ha	man
till	han	att_SN	kunna	ha
kompis	vi	ha	i	som

Table 6: 10 most frequent lemmas at CEFR levels A1-C1

We can see from the table that the most frequent word at A1 and A2 levels is the pronoun “jag” (Eng: “I”), which denotes that during the earlier levels, the student gradually learns how to talk about their daily lives and the people they associate with. This is also apparent from the most used nouns: “skola” (Eng: “school”) and “kompis” (Eng: “friend”). At level A2 we can see that more pronouns, “han” and “vi” (Eng: “he” and “we”, respectively), are included among the top ten words. This indicates that the learner is starting to refer to other people more frequently.

At the intermediate B levels, “jag” is no longer the top frequent word, but rather “vara” (Eng: “[to] be”). From this we can assume that the language in these levels becomes more about describing things and probably moves beyond the personal life prevalent at the A levels. Moreover, the verb “ha” (Eng: “have”) is introduced among the most frequent words at the B levels. In the Swedish language, “ha” is also used as an auxiliary verb in order to form the equivalent present and past perfect tenses. As such, the high frequencies of this word may be because the students are more acquainted with additional tenses.

An interesting addition to note at the C1 level is the presence of the lemma “som” (Eng: “who/which/as/that”). This is a clear indication that the student has reached a relatively proficient language level, being able to frequently construct subordinate clauses.

These are only a few examples, but they already show the students' language progress, through analysis of the most frequent words at this level. Our list gives the potential to the reader of this paper who is interested in language acquisition patters to explore further lexical patterns related to vocabulary progress.

8 Analysis

8.1 A Comparison between SweLL and SVALex

In this section, we compare the SweLL-list to SVALex. While SweLL-list consists of a productive type of vocabulary, depicted in learners' essays, SVALex contains a receptive type of vocabulary corresponding to the words that the learner is exposed to through the teaching material (including course books, vocabulary lists, dictionaries etc.). In order to conduct this comparison between the two lexical resources, we automatically matched lemma-pos combinations occurring in the SweLL-list with those in SVALex. We decided to take into consideration the POS tag of the entries when comparing the two resources for reasons of accuracy, so that for instance får_nn (Eng: “sheep”, noun) is not falsely compared to får_vb (Eng: “receives”, verb). We counted the entries in each of the lists, as well as the number of overlapping and missing entries. The comparison numbers for the lists are listed in Table 7.

Resource	**#items**	**#overlap**	**#missing**
SweLL	6,817	3,591	12,060
SVALex	15,681	3,591	3,226

Table 7: Comparison between SweLL-list and SVALex

As we can see in the table, SVALex is an extensive vocabulary list, almost twice the size of our SweLL-list. Consequently, it is not surprising that 12,060 entries present in SVALex are missing from the SweLL-list. On the other hand, there are 3,226 entries in the SweLL-list which are not present in SVALex. Although it would be interesting to have a closer look at these entries that are not present in SVALex, we chose to leave it out of this paper due to time constraints.

Furthermore, we found it interesting to compare the number of new lemmas per CEFR level that can be found in the SweLL-list and SVALex. In Table 8 we show the distribution of new lexical entries per CEFR level for both lists.

CEFR	**# of new items in SweLL**	**Norm. distribution in SweLL**	**# of new items in SVALex**	**Norm. distribution in SVALex**
A1	447	6.5%	1,157	7.4%
A2	1,326	19.4%	2,432	15.5%
B1	1,944	28.4%	4,332	27.7%
B2	1,222	17.9%	4,553	29.1%
C1	1,886	27.6%	3,160	20.2%

Table 8: Distribution of new entries per level in SweLL and SVALex

We can observe that the number of new entries for each level is higher in SVALex. This, we can assume, is happening because SVALex contains the receptive kind of knowledge, that which the learner

is taught during their lessons. In contrast, the SweLL-list includes the productive type of knowledge, which the learner actively uses in their essays.

In order to provide a proper comparison between the SweLL-list and SVALex, we also present a normalised distribution of new items per each level. This makes it easier to compare two corpora that have different sizes.

As the normalised distribution in SVALex shows, the number of new items to which the student is exposed to increases at level A2, and even more at the B levels. At C1, however, the increase in new entries is smaller than at the B levels. In the SweLL-list, the number of new items increases steadily from levels A1 until B1, but at level B2 the increase is much lower than at the previous two levels. This is a sharp contrast to SVALex, where B2 shows the highest increase in new words. Later, at C1 level, the relationship is reversed, with 20.2% increase in SVALex and an impressive 27.6% increase in SweLL. This is an interesting observation, and shows that while students learn a great part of their vocabulary at level B2, the productive vocabulary does not increase equally much until level C1. As such, this is a clear indicator that students are unable to actively use all their descriptive vocabulary and need some time to get used to using the words they have learned.

8.2 Analysis of the word level error normalisation

We mentioned in section 5 that we conducted a pre-normalisation procedure on the unlemmatised tokens, by first checking if we could find a match for these tokens in Saldom. After this procedure, we found that there were still 4,566 tokens that were not assigned a lemma. In order to solve this, we constructed an algorithm based on the Levenshtein distance, which would go through Saldom and choose the first lemma with the shortest possible LD to each unlemmatised token.

In this section, we present an analysis of the performance of our algorithm on the L2 word-level errors. We are not aware of any other project that have used the LD for normalisation of productive vocabulary essays and, as such, this is a test of the performance of LD in this context. We hope that this experiment can contribute towards improving the quality of essay annotation by eliminating L2 language deviations at the level of single words.

We proceeded with the evaluation of our normalisation algorithm by randomly selecting 20 of the unlemmatised tokens per CEFR level, and used our program to find the best match. Then, we checked if the returned lemma corresponded to the correct lemma, i.e. what the student intended to write (as far as our assumptions can go). We did this by analysing the context (sentence), into which we inserted the best matched lemma and observed whether it was semantically aligned.

We also did some quantitative analysis and have gathered some statistics on the performance of our algorithm in Table 9. Of the 20 lemmas we examined for each level, we show the number of correctly and incorrectly selected lemmas.

CEFR level	**Correctly returned lemma**
A1	7/20
A2	13/20
B1	13/20
B2	15/20
C1	16/20

Table 9: The performance of our Levenshtein distance based algorithm for 20 randomly selected tokens

As we can see from the results, the poorest performance of the error normalization occurs at level A1. However, it notably improves at the higher levels, reaching its best performance at level C1. In order to interpret these results, we needed to conduct a qualitative analysis and find the strong and weak points of using LD on L2 word level errors.

In Table 10 we have gathered a representative sample of sentences in which the misspelled entries occurred. We chose to include two sentences for each level, one which our algorithm found the correct match for, and one for which it did not.

CEFR	**Token**	**LD Match**	**Context**	**Correct**
A1	lagenhet	lägenhet	Jag bor i lägenhet	Yes
A1	monikor	monitor	Jag tycker om min monitor.	No (should be "människor")
A2	sekriva, naman	skriva, namn	Jag ska skriva sitt namn	Yes
A2	sister	sista	Jag bor i en lägenhet i N-gata tillsammans med min mamma, pappa och min lille bror och min två sista.	No (should be "systrar")
B1	ursprang	ursprung	Många säger att det finns cirka en tionde del människor som bor i Sverige och är inte svenskt ursprung.	Yes
B1	fran	fru	Människor fru olika kulturer har också påverkats hur svenskarna beteer sig.	No (should be "från")
B2	sammanfata	sammanfatta	För att kort sammanfatta jag tycker att det är ganska vktigt att prata om vilken roll stress spelar i våra liv.	Yes
B2	beretar	borsta	En undersökning i 2001 av Statistiska Centralbyrån borsta om att stressrelaterade och psykiska besvär har fördubblats sedan 1996.	No (should be "berättar")
C1	resetid	restid	Två timmars restid varje dag är inte onormal i Sverige.	Yes
C1	andå	and	Framför datorn men and hemma, kan man jobba för sitt företag.	No (should be "ändå")

Table 10: Examples of misspellings that were correctly and erroneously normalised

From this table we can see that our algorithm returns the right lemma in those cases where the student has written one letter erroneously. The LD normalisation correctly substitutes the misspelled letter, e.g.: lagenhet* → lägenhet, ursprang* → ursprung.

Furthermore, the algorithm seems to perform well when it can arrive at the correct lemma through omission of a single letter, e.g.: sekriva* → skriva, naman* → namn, resetid* → restid.

Also, in the case where a misspelled word lacks a single letter in order to form the correct word, the algorithm successfully adds the right letter and returns the correct lemma from SALDO, i.e. sammanfata* → sammanfatta.

On the other hand, when multiple misspellings occur in a word, the performance of our normalisation is rather poor, failing to fetch the correct lemma for the word which the student likely intended to write. Also, we can assume that in the case of the misspelled preposition “fran” (which should be “från”, Eng: “from”), the short length of the word plays a crucial part. Whenever a word is very short there will likely be many lemmas that have a Levenshtein distance of 1 from the token. Also, it is worth noting is that our Saldom list is alphabetically sorted, which means that lemmas that appear earlier in the alphabet are more likely to get picked. In the future, this could be substituted with a randomized version instead, or a more sophisticated additional filtering of candidates based on word co-occurrence measures.

We should also point out that in the case of the misspelling “andå”, it is the first letter that is misspelled. Since we needed to simplify our matching algorithm in order to make it perform reasonably fast, we chose to filter Saldom by only attempting to match lemmas starting with the same first letter as the token. Thus, our algorithm naturally failed to match this token correctly.

In conclusion, the Levenshtein distance seems to perform well when dealing with single letter errors which require a letter substitution, addition or omission. In contrast, its performance deteriorates when dealing with multiple errors on a word level. Also, short words can be more difficult to choose a correct match for, since there is a larger set of candidates to choose from. Our analysis also shows that Levenstein distance is applicable to normalization of writing at more advanced levels of language proficiency, whereas at the earlier stages it should be complemented by a more complex approach.

9 Conclusion and future work

Summing up our project, we started by digitising a number of essays, which were manually annotated in Lärka with metadata referencing the L2 students who wrote them, and their respective CEFR levels. The essays were then processed through the Korp pipeline, which performed linguistic annotation. Parsing the XML files into our Python program, we calculated lemma-pos frequencies, both raw and normalised, for all CEFR levels as well as for the list as a whole. For the tokens that Korp-pipeline was unable to fully annotate, we first applied a pre-normalisation procedure to find matching lemmas in Saldom. The remaining tokens that were still unlemmatised, to a great part misspellings, were then normalised using an algorithm based on the Levenshtein distance.

When our final list was created, we performed some analysis such as looking at the frequencies of new unique lemmas and examining the top 10 lemmas per CEFR level. We then proceeded to compare our productive vocabulary list to the receptive vocabulary list SVALex. Finally, we analysed the performance of our normalisation algorithm for L2 word errors.

In conclusion, we saw that a number of interesting observations could be made on our SweLL-list regarding the L2 learners' progression through the various CEFR levels. For example, we found that it is at the transition levels, from basic to intermediate and from the intermediate level to proficiency, that the highest vocabulary acquisition can be observed. Also, looking at the 10 most frequent lemmas for each level we could see a clear shift in the L2 learners' domain, from writing primarily about themselves to describing things, as well as improvement in their grammatical skills, with additional tenses being used and more complex sentences.

From our comparison with SVALex, we discovered that the overlap of lemmas were smaller than expected. Due to time constraints, we were unable to perform a proper analysis of why this was the case. However, we saw that while students learn a great part of their receptive vocabulary at level B2, the productive vocabulary does not increase equally much until level C1. This, we believe, indicates that students are unable to actively use all their receptive vocabulary and need some time to get used to using the receptive vocabulary they acquire at earlier levels.

The normalisation method we used was based on the Levenshtein distance. This turned out to perform reasonably well for single letter errors that required one substitution, addition or omission to find the target lemma. We found that to handle more efficiently short words, to which many lemmas would have a Levenshtein distance of 1, and for misspelled words with multiple errors we found our normalisation algorithm to be less effective.

We feel that there are many more interesting tasks that can be done to improve the quality of L2 learners' essay annotation. Of these, we saw especially that using the Levenshtein distance for normalisation has potential, but it might need to be combined with other NLP tools or methods in order to improve its performance. For example, instead of simply choosing the first lemma with the closest distance, we could create a set of heuristics to decide among the equally likely candidates. For longer words, we could also analyse the tokens to see if compounding is the reason they have not been assigned a lemma. Other versions of LD, such as Damerau-Levenshtein (see e.g. Cucerzan & Brill, 2004) should also be considered.

One suggestion for future work is to upload the SweLL-list to the same visualisation engine as SVALex. This would facilitate easier exploration the SweLL-list by itself as well as further comparison with SVALex. Note, however, that their respective frequencies are not fully comparable since SVALex uses the distribution index, while SweLL uses WPM. Furthermore, while we were able to make interesting observations from our SweLL-list and its comparison to SVALex, the limited size and domain of the list means that our findings may be inconclusive for a general assessment of L2

learners' skills and progression in the current stage. There were, however, several indications that could be confirmed if the SweLL corpus continues to grow.

10 References

L. Antonsen. Improving Feedback on L2 Misspellings - An FST Approach. In *Proceedings of the SLTC 2012 workshop on NLP for CALL*. Linköping Electronic Conference Proceedings, 2012.

S. Bird. NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions* (pp. 69-72) 2006. Association for Computational Linguistics.

Lars Borin, Markus Forsberg, and Johan Roxendal. Korp – The Corpus Infrastructure of Språkbanken. In *Proceedings of LREC 2012. Istanbul: ELRA*, page 474–478, 2012. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/248_Paper.pdf

Lars Borin, Markus Forsberg, and Lennart Lönngren. Saldo: A touch of Yin to Wordnet's Yang. *Language resources and evaluation*, 47(4):1191–1211, 2013. URL http://dx.doi.org/10. 1007/s10579-013-9233-4

Annette Capel. A1–B2 Vocabulary: Insights and Issues Arising from the English profile Wordlists project. *English Profile Journal*, 1:e3 (11 pages), 9 2010. ISSN 2041-5362. URL http: //journals.cambridge.org/article_S2041536210000048

Frieda Charalabopoulou, Maria Gavrilidou, Sofie Johansson Kokkinakis, and Elena Volodina. Building Corpus-Informed Word Lists for L2 Vocabulary Learning in Nine Languages. In *CALL: Using, Learning, Knowing. EuroCALL Conference, Gothenburg, Sweden, 22-25 August 2012, Proceedings. Eds. Linda Bradley and Sylvie Thouësny. Research-publishing.net, Dublin, Ireland*, volume 2012, 2012. ISBN 978-1-908416-03- 2. URL http://research-publishing.net/publication/chapters/978-1908416-03-2/ Charalabopoulou_Gavrilidou_et_al_25.pdf .

S. Cucerzan and E. Brill. Spelling Correction as an Iterative Process that Exploits the Collective Knowledge of Web Users. In *Proceedings of EMNLP 2004*, pages 293–300, July 2004.

Rachele De Felice and Stephen G. Pulman. A Classifier-Based Approach to Preposition and Determiner Error Correction in L2 english. In *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, COLING '08, pages 169–176, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics. ISBN 978-1-905593-44-6. URL http://dl.acm.org/citation.cfm?id=1599081.1599103

Eva Forsbom. Deriving a Base Vocabulary Pool from the Stockholm-Umeå Corpus, 2006.

Thomas François and Cédrick Fairon. An "Ai Readability" Formula for French as a Foreign Language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 466–477, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?Id=2390948.2391004.

Thomas François, Nuria Gala, Patrick Watrin, and Cédrick Fairon. Flelex: a Graded Lexical Resource for French Foreign Learners. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014*., pages 3766–3773, 2014. URL http://www.lrec-conf.org/proceedings/lrec2014/ summaries/1108.html.

Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. Svalex: a CEFR-Graded Lexical Resource for swedish foreign and second language learners. To appear in *Proceedings of LREC 2016*, Slovenia.

A. W. Hauser and K. U. Schulz. Unsupervised Learning of Edit Distance Weights for Retrieving Historical Spelling Variations. In *Proceedings of the First Workshop on Finite-State Techniques and Approximate Search*, 2007.

Katarina Heimann Mühlenbock and Göteborgs universitet. Institutionen för svenska språket. I See What You Mean: Assessing Readability for Specific Target Groups, 2013.

Ann-Kristin Hult. Old and New User Study Methods Combined - Linking Web Questionnaires with Log Files from the Swedish Lexin Dictionary. In Ruth Vatvedt Fjeld and Julie Matilde Torjusen, editors, *Proceedings of the 15th EURALEX International Congress*, pages 922–928, Oslo,Norway, aug 2012. Department of Linguistics and Scandinavian Studies, University of Oslo. ISBN 978-82-303-2228-4.

Håkan Jansson, Sofie Johansson Kokkinakis, Judy Carola Ribeck, and Emma Sköldberg. A Swedish Academic Word List: Methods and Data. In *Proceedings of the 15th EURALEX International Congress 7-11 August, 2012, Oslo*, pages 955–960, 2012. ISBN 978-82-303-2228-4.
URL http://www.euralex.org/proceedings-toc/euralex_2012/.

VI Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707, 1966.

I.S.P. Nation. *Learning Vocabulary in Another Language*. Cambridge Applied Linguistics. Cambridge University Press, 2001. ISBN 9780521800921.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. The Conll-2014 Shared Task on Grammatical Error Correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*: Shared Task, pages 1–14 ,Baltimore, Maryland, June 2014. Association for Computational Linguistics.
URL http://www.aclweb.org/anthology/W/W14/W14-1701

Council of Europe and Council for Cultural Co-operation. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, 2001.

Ildikó Pilán, Elena Volodina, and Richard Johansson. Rule-Based and Machine Learning Approaches for Second Language Sentence-Level Readability. In *Proceeding of the ACL 2014 9th Workshop on Innovative Use of NLP for Building Educational Applications, Baltimore, june 22-27 2014*, pages 174–184, 2014. ISBN 978-1-941643-03-7.
URL http: //www.aclweb.org/anthology/W/W14/W14- 1821.pdf .

English Profile. Introducing the CEFR for English, 2011.
URL http://www.englishprofile.org/images/pdf/theenglishprofilebooklet.pdf .

Anne Rimrott and Trude Heift. Language Learners and Generic Spell Checkers in Call – Calico journal. *CALICO*, 23(1), September 2005.

Elena Volodina and Sofie Johansson Kokkinakis. Introducing Swedish Kelly-List, a New Free E-resource for Swedish. In *LREC 2012 Proceedings*, volume 2012, 2012a.
URL http://www. lrec-conf.org/proceedings/lrec2012/pdf/264_Paper.pdf

Elena Volodina and Sofie Johansson Kokkinakis. Swedish Kelly: Technical Report. Technical Report, Department of Swedish, 2012b.
URL https://gupea.ub.gu.se/bitstream/2077/28860/1/gupea_2077_28860_1.pdf .

Elena Volodina and Therese Lindström TIedemann. Evaluating Students' Metalinguistic Knowledge with Lärka., 2014.
URL http://spraakbanken.gu.se/sites/spraakbanken.gu.se/files/sltc2014_submission_35_filan_v2.pdf

Elena Volodina, Ildikó Pilán, Stian Rødven Eide, and Hannes Heidarsson. You Get What You Annotate: a Pedagogically Annotated Corpus of Coursebooks for Swedish as a Second Language. In *NEALT Proceedings Series*, volume 22, pages 128–144, 2014. ISBN 978-91-7519-175-1.
URL http://www.ep.liu.se/ecp/107/010/ecp14107010.pdf .

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. Swell on the Rise: Swedish Learner Language Corpus for European Reference Level Studies. To appear in *Proceedings of LREC 2016*, Slovenia