# Towards Single Word Lexical Complexity Prediction

**David Alfter**
Språkbanken
University of Gothenburg
Sweden
david.alfter@gu.se

**Elena Volodina**
Språkbanken
University of Gothenburg
Sweden
elena.volodina@gu.se

## Abstract

In this paper we present work-in-progress where we investigate the usefulness of previously created word lists to the task of single-word lexical complexity analysis and prediction of the complexity level for learners of Swedish as a second language. The word lists used map each word to a single CEFR level, and the task consists of predicting CEFR levels for unseen words. In contrast to previous work on word-level lexical complexity, we experiment with topics as additional features and show that linking words to topics significantly increases accuracy of classification.

## 1 Introduction

A way of addressing the second-language (L2) acquisition needs of the recent influx of new immigrants to Sweden would be to provide an extensive amount of digitally accessible self-study materials for practice. This could be achieved through the development of specific algorithms for exercise/material generation, but such algorithms generally heavily rely on linguistic resources, such as descriptions of vocabulary and grammar scopes per each stage of language development, so that automatic generation of learning materials would follow some order of increasing complexity.

Vocabulary scope can be described through graded vocabulary lists. These are lexical resources where each lexical item is linked to a level at which the item is appropriate for learners to study, one prominent example being the English Vocabulary Profile (Capel, 2010, 2012). Graded lexical resources are useful, for example, for course book writers, language test designers, language teachers and language learners, since they can inform the users as to what knowledge is to be expected at which proficiency level, as well as which words to teach and test at which levels.

However, any graded list is a finite resource, as it would never be possible to list by levels all items that learners might encounter. We intend, therefore, to use previously compiled graded vocabulary lists to learn from them to predict levels of previously unseen, out-of-vocabulary (OOV), lexical items.

In practical terms, we look at three automatically created corpus-based vocabulary lists, namely Kelly list (Volodina and Kokkinakis, 2012), a resource based on L1 web corpora that identifies frequent vocabulary to guide language learners in their acquisition of vocabulary[1], as well as SVALex (François et al., 2016) and SweLLex (Volodina et al., 2016b), two L2-targeted word lists covering receptive vocabulary and productive vocabulary respectively[2]. The aim of this work is, thus, to create a model that is able to predict the difficulty (i.e. appropriate CEFR[3] level) of any Swedish word with regard to productive and receptive aspects. These graded vocabulary lists are then intended for use in generation of exercises for learners of different levels, though other usage scenarios are also possible.

## 2 Related Work

There has been some work on the creation and evaluation of automatically graded vocabulary lists (Gala et al., 2013, 2014; Tack et al., 2016b).

Gala et al. (2013) aim at identifying criteria that make words easy to understand, independently of the context in which they appear. Since it has been shown that the concept of difficulty depends on the target group (Blache, 2011; François,

[1] Swedish Kelly list is available with CC-BY license from *https://spraakbanken.gu.se/eng/resource/kelly*

[2] Both lists are a part of CEFRLex family of resources, and are available from *http://cental.uclouvain.be/cefrlex/*

[3] Common European Framework of Reference for Languages (Council of Europe, 2001) describes six levels of proficiency, starting from A1 to C2

*Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 79–88
New Orleans, Louisiana, June 5, 2018. ©2018 Association for Computational Linguistics

2012), and thus different combinations of features might model certain groups better than others, they focus on speech productions by patients with Parkinson's disease. Gala et al. (2013) look at 27 intra-lexical and psycholinguistic variables. The intra-lexical variables include number of letters, number of phonemes, number of syllables, syllable structure (CV structure), consistency between graphemes and phonemes, and selected difficult spelling patterns such as double vowels and double consonants. Among psycholinguistic variables are orthographic neighborhood (words that only differ by one letter), lexical frequency and presence/absence from the Gougenheim list, a list of easy-to-understand vocabulary items.

They train a Support Vector Machine (SVM) classifier on the nine (out of initial 27) most predictive features to predict the difficulty level of unseen words. 5-fold cross-validation on the data shows an average accuracy of 62% in the three-way classification. They conclude that syllabic structures and spelling patterns are not very predictive of difficulty and that the most predictive features are the lexical frequency and presence/absence from the Gougenheim list.

Gala et al. (2014) focus on learners of French, both L1 learners and learners of French as a foreign language. They use Manulex (Lété et al., 2004) to model L1 learners' vocabulary and FLELex (François et al., 2014) to model L2 learners' vocabulary. In contrast to Gala et al. (2013), they use 49 features which can be grouped into orthographic features (e.g. number of letters, number of phonemes, number of syllables), morphological features (number of morphemes, affix frequency, compounding), semantic features (degree of polysemy) and statistical features (frequency, Gougenheim list). They train two SVM classifiers, one for L1 learners and one for learners of French as a foreign language. The first one is a three-way classification while the latter is a six-way classification. On the three-way classification, they reach 63% accuracy and on the six-way classification they reach 43% accuracy. As in Gala et al. (2013), they find the most predictive features to be lexical frequency and presence/absence from the Gougenheim list. However, they also find the binary polysemous status, i.e. whether the word polysemous or not, as well as the degree of polysemy to correlate well with the complexity of words. This is an interesting finding, as the degree of polysemy is not directly correlated with frequency.

A related area of work is complex word identification for text simplification. For this task, it is important to identify target *difficult* words or phrases that need simplification (Shardlow, 2013; Paetzold and Specia, 2016; Štajner et al., 2018). However, in contrast to our work, complex word identification is a binary classification and the focus is slightly different, although there are significant overlaps. Tack et al. (2016a) and Tack et al. (2016b) for example aim at identifying and classifying words of a text into known and unknown ones either for an individual learner or for learners of a given proficiency level as a group. They compare different personalized models with a model based on the graded vocabulary list FLELex (François et al., 2014). Their personalized models also use frequency information, CEFR levels of single words as calculated in Gala et al. (2014), number of letters, and number of senses of a word. For the FLELex vocabulary based model and a learner of a given CEFR level, the model considers all words that are of the same or lower level as the learner's level as known and all words that are of higher level as unknown.

Our recent participation in the Complex Word Identification Task 2018 (Štajner et al., 2018) has yielded interesting findings that we hope will further improve the presented system (Alfter and Pilán, 2018).

## 3 Data

Our data consists of three different word lists for Swedish, namely SVALex (François et al., 2016), SweLLex (Volodina et al., 2016b) and Kelly list (Volodina and Kokkinakis, 2012).

SVALex is compiled from the COCTAILL textbook corpus (Volodina et al., 2014), comprised of reading comprehension texts marked for CEFR levels, and covers receptive vocabulary knowledge. SweLLex is derived from the pilot SweLL learner essay corpus (Volodina et al., 2016a) graded for CEFR levels and covers productive vocabulary knowledge. Kelly list is derived from the Swedish Web-as-Corpus (SweWaC) and contains the 8425 most frequent lemmas appearing in native speaker writing divided into CEFR level according to the frequency of the items and corpus coverage. See table 1 for the overview of the three resources.

While Kelly list already assigns each word to a

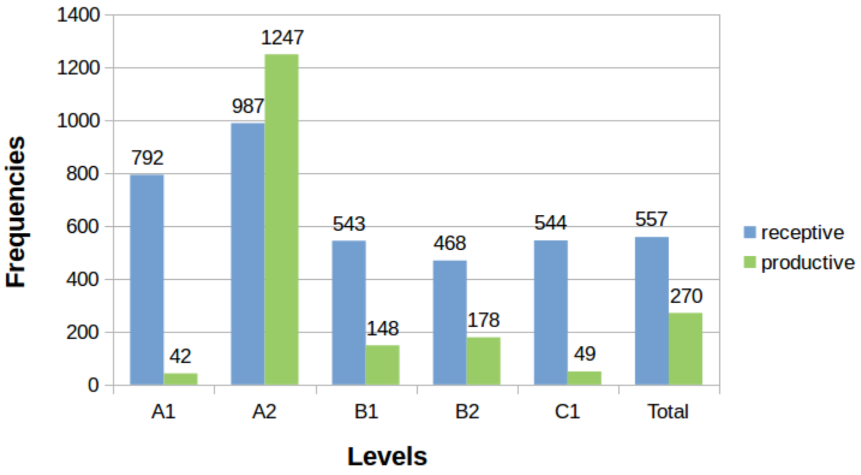


Figure 1: Distribution of the verb *arbeta* ‘to work’, in receptive and productive resources

| | A1 | A2 | B1 | B2 | C1 | Total |
|---|---|---|---|---|---|---|
| SVALex | 968 | 1973 | 2761 | 6223 | 3697 | 15 681 |
| SweLLex | 602 | 1258 | 1317 | 1024 | 1248 | 6 965 |
| Kelly list | 1404 | 1404 | 1404 | 1404 | 2809 | 8 425 |

Table 1: Data distribution across lists. In SVALex and SweLLex vocabulary items partially overlap between levels, and hence the total number of items in the list does not equal the sum of items per level.

target CEFR level, SVALex and SweLLex present distributions over CEFR levels, i.e. how often a word occurs at the different CEFR levels, as exemplified in table 2. Since SVALex and SweLLex cover 5 proficiency levels and Kelly list covers 6 proficiency levels, we assimilated the highest level in Kelly list (C2) to the previous level (C1).

To go from distributions to target levels in SweLLex and SVALex, we use the mapping procedures described in Gala et al. (2013), Gala et al. (2014) (first occurrence) and Alfter et al. (2016) (threshold). For *first-occurrence mapping*, we assign each word to the level it first occurs at. For *threshold mapping*, we assign each word to the level where it occurs *significantly* more often than at the preceding level, with the level of significance set at 30%.

Figure 1 shows the distribution of frequencies for the word *arbeta* (Eng. “to work”) over the five CEFR levels in SVALex (receptive resource, 1st bar) and SweLLex (productive resource, 2nd bar). According to the *first occurrence* approach, the target level for both receptive and productive competence for the word *arbeta* would be A1, whereas the *threshold* approach suggests that A1 would be the target level for receptive knowledge, and A2 would be the target level for productive level.

We did a comparison of both mapping methods to find out to what degree they agree. Table 3 shows the levels assigned by both methods for the two resources SVALex and SweLLex. By comparing the output of these two mapping methods, we can see that both methods agree to a large extend. When both methods did not agree, they tended to still assign levels that were adjacent, e.g. if one method assigned level B1, the other would assign B2 or A2. This is not a surprise, as the border between different proficiency levels can be fluid. We call this type of disagreement *within one level*. We also see that a certain amount of words were classified as different levels but with the levels assigned being more than one level apart, e.g. one method assigns level A2 and the other method assigns level B2. We call this type of disagreement *more than one level*. Given this finding, and for comparability between studies, e.g. with Gala et al. (2013) and Gala et al. (2014), we have opted to use the first-occurrence approach in the remainder of the study.

The SVALex and SweLLex data is noisy, because, for one, we cannot validate whether the automatically assigned (mapped) levels are accurate

| Lemma | Part-of-Speech | A1 | A2 | B1 | B2 | C1 |
|---|---|---|---|---|---|---|
| beta ‘to graze’ | VB | 0.0 | 0.0 | 0.0 | 19.27 | 13.21 |
| bo ‘to live’ | VB | 4978.93 | 2515.92 | 1252.19 | 718.53 | 497.75 |
| hund ‘dog’ | NN | 251.89 | 81.26 | 250.26 | 74.29 | 98.87 |

Table 2: Example of word distributions over levels in SVALex

| Resource | Same level | Within one level | More than one level |
|---|---|---|---|
| SVALex | 12775 | 1592 | 1255 |
| SweLLex | 5689 | 706 | 516 |

Table 3: Number of items that were assigned the same level, within one level and more than level by both mapping techniques

due to missing gold standard annotations, and secondly because of certain errors resulting from automatic corpus annotation. The data is also sparse, and since the mapping procedure for SVALex and SweLLex very much depends on the data available, this introduces further noise. These are the limitations we are aware of and plan to address in the future by collecting and annotating more data.

## 4 Features

From each word, including multi-word expressions such as *göra ont* ‘to hurt’ and *god morgon* ‘good morning’, we extract features, grouped into count-based features (i), morphological features (ii), semantic features (iii) and context-based features (iv). Table 4 gives an overview of the average values for some selected features per level and resource. As can be seen from this table, words at higher levels tend to be longer, have more syllables, longer suffixes, a higher number of compounds and lower degrees of polysemy and homonymy. Indeed, concerning polysemy, more common words, which are typically found at lower levels, tend to have more different senses than more specialized words found at higher levels.

### (i) Count-based and surface form features

- *Length* is the length of the word in characters, our example word *arbeta* (Eng “to work”) containing 6 characters. Word length has previously been used to assess linguistic complexity, among others in readability assessment formulas, for example in Smith (1961); Björnsson (1968); O’Regan and Jacobs (1992).

- *Syllable* count is the number of syllables in the word, where *arbeta* contains three syllables. Syllables are counted as number of vowels except for diphthongs ending in ‘u’ (e.g. ‘eu’, ‘au’) which are counted as one syllable. Syllable count has been applied in readability assessment as a measure of increasing text difficulty, e.g. in Flesch (1948); Kincaid et al. (1975), where multi-syllable words have been proven to increase the overall linguistic complexity of a text. By analogy, we assume that the same applies on a single word level.

- *Contains non-alphanumeric characters* is a boolean value that is true if the word contains non-alphanumeric characters, i.e. any character other than A-Z and digits 0-9, for example *13-åring* (Eng. 13-year old).

- *Contains number* is a boolean value that is true if the word contains digits or consists solely of digits.

- The *multi-word* feature indicates whether the lexical expression is made up of more than one single word.

- For *bigrams*, we calculated all character-level bigrams from each word list and retained only the 53 most predictive ones. This feature is a vector indicating the presence or absence of these 53 bigrams in the target word.

- For *n-gram probabilities*, we calculate character-level unigram, bigram and trigram probabilities with a language model based on the Swedish Wikipedia dump from February 2018. We surmise this also implicitly cap-

| | A1 | A2 | B1 | B2 | C1 |
|---|---|---|---|---|---|
| **Average word length** | | | | | |
| SVALex | 6.00 | 7.49 | 8.51 | 8.85 | 9.58 |
| SweLLex | 5.10 | 5.98 | 7.66 | 8.89 | 9.91 |
| Kelly | 5.74 | 7.00 | 7.54 | 7.86 | 7.80 |
| **Average syllable count** | | | | | |
| SVALex | 2.08 | 2.52 | 2.88 | 2.91 | 3.24 |
| SweLLex | 1.80 | 2.01 | 2.58 | 2.94 | 3.28 |
| Kelly | 2.04 | 2.44 | 2.62 | 2.78 | 2.76 |
| **Average suffix length** | | | | | |
| SVALex | 0.54 | 0.63 | 0.77 | 0.80 | 0.91 |
| SweLLex | 0.47 | 0.51 | 0.56 | 0.63 | 0.71 |
| Kelly | 0.70 | 0.80 | 0.86 | 0.88 | 0.87 |
| **Average number of compounds** | | | | | |
| SVALex | 0.014 | 0.037 | 0.052 | 0.062 | 0.067 |
| SweLLex | 0.038 | 0.058 | 0.112 | 0.125 | 0.162 |
| Kelly | 0.043 | 0.095 | 0.137 | 0.175 | 0.167 |
| **Average degree of polysemy** | | | | | |
| SVALex | 0.64 | 0.51 | 0.39 | 0.29 | 0.24 |
| SweLLex | 0.55 | 0.62 | 0.46 | 0.36 | 0.30 |
| Kelly | 0.84 | 0.73 | 0.67 | 0.56 | 0.56 |
| **Average degree of homonymy** | | | | | |
| SVALex | 1.25 | 1.11 | 1.06 | 1.05 | 1.02 |
| SweLLex | 1.35 | 1.18 | 1.10 | 1.08 | 1.04 |
| Kelly | 1.30 | 1.13 | 1.08 | 1.10 | 1.05 |

Table 4: (Selected) feature averages per level and resource

tures information about grapheme-phoneme correspondence, frequency and suffixes.

**(ii) Morphological features**

- *Part-of-speech* corresponds to the part-of-speech of the word. For multi-word expressions, the part-of-speech of the head noun is taken.

- For *suffix length*, we stem the word using the NLTK stemmer (Bird et al., 2009) and subtract the length of the resulting stem from the length of the original word. In *arbeta*, the final -a is a suffix. Previous work on order of acquisition of inflectional versus derivational morphemes, e.g. Derwing (1976), argue that knowledge of derivational morphology is acquired gradually in the learning progress, thus motivating this feature for our experiments. This intuition also seems to hold when looking at average suffix length by level, as shown in table 4.

- For *compound count*, we run the word through the SPyRo/SALDO pipeline (Östling and Wirén, 2013), which generates possible analyses of the word with regard to compounding. Compound count is the number of possible compounding alternatives. *Arbeta* can theoretically be analyzed as *ar* 'are (unit of measurement)' + *beta* 'to graze' and thus would have a compound count of 1. *Glasskål* on the other hand can be analyzed as *glas* 'glass' + *skål* 'bowl', *glass* 'ice cream' + *skål* 'bowl' and *glass* 'ice cream' + *kål* 'cabbage' and thus would have a compound count of 3. The cognitive load for processing a word, that potentially has several (compounding) interpretations, hypothetically also influences the word's complexity, and hence the level at which it is acquired.

- For *compounds*, we calculate all compound elements, i.e. words that have been identified in compounds, in all lists and selected the 12 most predictive compounds. This feature is a vector indicating the presence or absence of these compounds in the target word.

- *Gender* for nouns is taken from Saldo's morphology (Borin et al., 2008) and encoded numerically as -1 (no information about gender or not applicable), 0 (common gender, aka "en-ord"), 1 (neuter, aka "ett-ord") and 2 (variable gender). For *arbeta* the value would be -1 since gender only applies to nouns. The majority of nouns in Swedish are of common gender (e.g. in the Kelly-list there are 3465 nouns of common gender, while 1065 are neuter).

**(iii) Semantic features**

- *Degree of polysemy* is calculated by counting the sub-entries of a given dictionary entry in Lexin (Gellerstam, 1999). The verb *arbeta* has only one sub-entry, and is thus non-polysemous. From empirical sources (e.g. various frequency lists), we can observe that non-polysemous words tend to be less used constituting a large bulk of non-frequent words, something that is quite logical given that most word lists are compiled based on lem-grams (e.g. a combination of base form of a word plus its part-of-speech), and not on senses. Usages of several senses of the same lem-gram are thus grouped together in one entry and push the word to the top of the frequency lists. Highly polysemous words, like *komma* 'to come' are thus often learned in the beginning. This seems to be a contradictory trend with regards to our example word, *arbeta* 'to work'. However, if we extend the search to phrasal verbs with *arbeta* in Saldo, there would be seven more entries, and in Lexin four more.

- *Degree of homonymy* is calculated by counting the number of dictionary entries in Lexin with the same orthographic form. An example of a homonym across word classes would be *gift*: it could either be the adjective meaning "married" or the noun meaning "poison". Homonymy within the same word class would be *vara* (Eng. "to last", "to be"). The example word *arbeta* has only one entry in Lexin. Studies on homonymy within second language learning (Mashhady et al., 2012) show that honomymous words take longer to remember and differentiate between meanings than e.g. several synonyms relating to the same concept, demanding disambiguation of a homonym given the context, which makes homonymy an interesting feature to include into our experiments.

**(iv) Context features**

- For *topic distributions*, we indicate in which topic lists the target word occurs. Topic lists were extracted from the COCTAILL corpus, where each reading text is assigned one or more topics. We thus extracted all lemmata from reading texts, assigning them to the topics as given in the corpus. We then ran a TF-IDF algorithm over the lists to eliminate words that occurred across all topic lists. This yielded 33 topic lists, such as animals, arts, daily life, food and drink, nature, places, or technology.

Thus, for the verb *arbeta*, we can summarize the above features into the following (simplified) word complexity description: 6-letter 3-syllable non-polysemous non-homonymous verb with one possible suffix, one possible compound analysis, no gender information (since this only applies to nouns), not a multi-word expression and a word used in topics characteristic of presenting people (CEFR levels A1 and A2) which is - supposedly - the reason why the empiric data points out A1 level for receptive and productive knowledge according to *first-occurrence* approach; and A1 for receptive and A2 for productive knowledge if we follow the *threshold* mapping strategy.

## 5 Classification

In order to check how well the features we have chosen model single word complexity, we use different classifiers and stratified 10-fold cross-validation on the different data sets.

For classification of unseen words, we train classifiers on the available data. We train one classifier for receptive predictions on SVALex and one classifier for productive predictions on SweLLex.

The classification task consists in assigning each word in our word lists a target CEFR level.

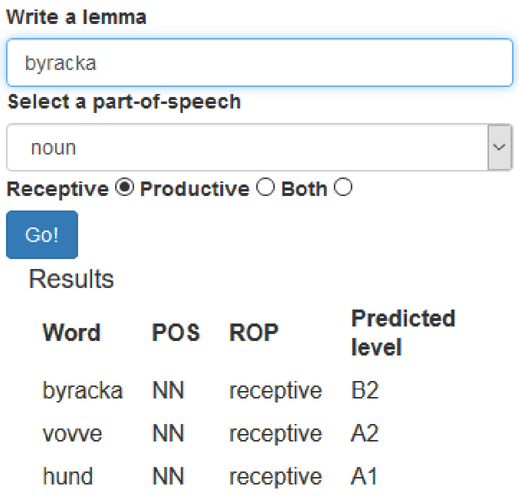
|  | Svalex | Swellex | Kelly |
|---|---|---|---|
| Majority baseline | 0.29 ± 0.00 | 0.29 ± 0.00 | 0.33 ± 0.00 |
| SVM | 0.32 ± 0.02 | 0.37 ± 0.05 | 0.39 ± 0.04 |
| MLP | 0.32 ± 0.03 | 0.37 ± 0.04 | 0.39 ± 0.04 |
| ET | 0.27 ± 0.02 | 0.33 ± 0.05 | 0.32 ± 0.04 |
| SVM+T | 0.44 ± 0.03 | **0.41** ± 0.04 | **0.45** ± 0.05 |
| MLP+T | 0.53 ± 0.04 | 0.38 ± 0.05 | 0.44 ± 0.05 |
| ET+T | **0.55** ± 0.05 | 0.37 ± 0.06 | 0.43 ± 0.05 |
| SVM+TL | 0.48 ± 0.03 | 0.41 ± 0.05 | 0.45 ± 0.04 |
| MLP+TL | 0.53 ± 0.04 | 0.39 ± 0.06 | 0.44 ± 0.03 |
| ET+TL | **0.59** ± 0.03 | 0.37 ± 0.06 | 0.42 ± 0.03 |

Table 5: Results: Accuracy and standard deviation using 10-fold cross-validation

For evaluation of the features, accuracy is calculated by comparing the predicted level with the level given by the graded word list. We cannot, at this moment, evaluate classifiers for unseen words, as we would have to have manually graded word lists against which to compare our predictions.

## 6 Results

Table 5 shows the results of 10-fold cross-validation classification using different algorithms. Majority baseline always predicts the majority class. Since our data is not balanced, this deviates from the expected chance baseline of 0.2 for five-class classification. SVM is a support vector machine with default parameters $C = 1$ and radial basis function (rbf) kernel. MLP is a multilayer perceptron with 100 hidden layers and a learning rate of 0.01. These parameters were chosen based on a randomized grid search over the parameter space. ET is an extra trees classifier, a classifier from the group of random tree classifiers. Preliminary experiments have shown an initial increase in accuracy with an increase in the number of estimators of the ET algorithm but which shows no further improvement after 100 estimators. We thus have fixed the number of estimators for the ET algorithm at 100. SVM+T, MLP+T and ET+T show the accuracies obtained by the same algorithms but with topic distributions added to the data. For comparability, since we have included all word classes in our experiments, we also tried classifying only lexical word classes (nouns, verbs, adjectives and adverbs) as in Gala et al. (2014). The results of these experiments are shown in the rows SVM+TL, MLP+TL and ET+TL.



Write a lemma

byracka

Select a part-of-speech

noun

Receptive ● Productive ○ Both ○

Go!

Results

| Word | POS | ROP | Predicted level |
|---|---|---|---|
| byracka | NN | receptive | B2 |
| vovve | NN | receptive | A2 |
| hund | NN | receptive | A1 |

Figure 2: User interface for lexical complexity prediction

In addition, we have created a user interface[4], as shown in figure 2. This user interface can be used for getting predictions of any word, not only words present in the word lists . The input word is transformed into a feature vector as described above and then fed into the classifier, which predicts a label. Figure 2 shows the predictions for *hund* 'dog', *vovve* 'childish or endearing term for dog' and *byracka* 'derogatory term for dog'.

## 7 Discussion

We found that our features excluding topic distributions barely outperform the majority baseline, yielding even lower scores than the baseline in some cases. Adding topic distributions signifi-

[4] https://spraakbanken.gu.se/larkalabb/siwoco

cantly improves accuracy.

In comparison to the results presented in Gala et al. (2014), we can see an expected trend. Indeed, on the L1 resource Manulex and Kelly (which is based on L1 data but intended for L2 audiences), they reach 63% accuracy in a three-way classification while we reach 45% accuracy in a five-way classification. On the L2 textbook corpus resources FLELex and SVALex, they reach 43% accuracy in a six-way classification while we reach 59% accuracy in a five-way classification.

If we are comparing our results without topic distributions, which are more similar to the results presented in Gala et al. (2014) due to the similarity of features, we see that our best system on L2 data performs worse in a five-way classification (0.32) than theirs in a six-way classification (0.43). This is probably due to the size of the corpus that was used to compile these lists. While FLELex was compiled from 28 textbooks and 29 readers, COCTAILL was compiled from 12 textbooks only. As such, their distributions are less sparse and hypotheses about the target level can be made with more certainty.

Another point is that, in contrast to previous work, we have not included information about lexical frequency explicitly. Including such information could possibly further improve accuracy. It can be argued that n-gram probabilities latently encode this information, but it would be interesting to see whether a more explicit approach would lead to better results.

We also ran cross-validated recursive feature elimination (Guyon et al., 2002) to get a ranking of features and discard useless features. This interestingly identified bigram features (presence/absence of most predictive bigrams; not to be confused with bigram frequency) and compound features as useless, but excluding those features does not lead to an increase in accuracy. However, looking at the most predictive bigram and compound files, it seems that something went wrong during calculation of these, since, for example in bigrams, there are only very rare combinations such as ‘åä’, ‘åo’, ‘xf’ and ‘xb’. We would like to address this issue in future work. The final model uses 64 features.

One problem for the classifiers could be that representing words as vectors can lead to the same representation for different words with different levels, which leads to a decrease in learnability since it introduces contradictory data points. We have checked for this and found out that our data contains about 5% of contradictory data points. A possible approach could be to add more disambiguating features.

## 8 Conclusion and future work

We have presented insights from work-in-progress on single word lexical complexity. In contrast to previous work, we show that adding topic information significantly improves results on the classification task. However, the current topic lists can be further refined, for example by synonym expansion, in the hope of improving accuracy.

For future work, one concern that was also expressed in Gala et al. (2014) is that the current lists do not discriminate between different senses of a word. Thus, words like *glas*, meaning either ‘glass’ as substance or ‘glass’ as receptacle for drinks, would be assigned one single level while their different senses clearly should be assigned different levels. We are currently working on recalculating the resources SVALex and SweLLex on the sense level by including a word sense disambiguation component in the pipeline.

Another interesting experiment could be to include number of phonemes in our study, since Swedish has some non-transparent grapheme-to-phoneme correspondences.

There is currently ongoing work concerning the collection and annotation of learner essays, which we hope will alleviate the data sparseness problem that we face at the moment, especially with regard to the learner essay based word list.

We would also like to implicitly crowdsource learner knowledge by embedding words from these automatically mapped lists in automatically generated learner exercises. By monitoring how learners of a given level are dealing with words predicted to be of their level, we hope to be able to draw conclusions about the target level of words, i.e. if learners of intermediate B1 level consistently have problems with certain words that our mapping predicts to be of B1 level, we can assume that the prediction was incorrect.

In the future, we intend to evaluate these resources both with teachers of Swedish as a second language as well as language learners to estimate the validity of the automatic mapping. We would also like to create gold standard annotations, both based on these resources as well as new resources.

## 9 Acknowledgements

This work has in part been funded by an infrastructure grant from the Swedish Research Council to Swedish National Language Bank. We would also like to thank the anonymous reviewers for their constructive feedback.

## References

David Alfter, Yuri Bizzoni, Anders Agebjörn, Elena Volodina, and Ildikó Pilán. 2016. From Distributions to Labels: A Lexical Proficiency Analysis using Learner Corpora. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016*, 130, pages 1–7. Linköping University Electronic Press.

David Alfter and Ildikó Pilán. 2018. SB@GU at the Complex Word Identification Task 2018. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Carl Hugo Björnsson. 1968. *Läsbarhet*. Liber.

Philippe Blache. 2011. A computational model for linguistic complexity.

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2008. The hunting of the BLARK–SALDO, a freely available lexical database for Swedish language technology. *Resourceful language technology. Festschrift in honor of Anna Sågvall Hein*, (7):21–32.

Annette Capel. 2010. A1–B2 vocabulary: insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, 1(1):1–11.

Annette Capel. 2012. Completing the English Vocabulary Profile: C1 and C2 vocabulary. *English Profile Journal*, 3:1–14.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

Bruce L Derwing. 1976. Morpheme recognition and the learning of rules for derivational morphology 1. *Canadian Journal of Linguistics/Revue canadienne de linguistique*, 21(1):38–66.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Thomas François. 2012. *Lexical and syntactic complexities: a difficulty model for automatic generation of language exercises in FFL*. Ph.D. thesis, Université Catholique de Louvain, Louvain-la-Neuve.

Thomas François, Núria Gala, Patrick Watrin, and Cédrick Fairon. 2014. Flelex: a graded lexical resource for french foreign learners. In *LREC*, pages 3766–3773.

Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: a CEFR-graded Lexical Resource for Swedish Foreign and Second Language Learners. In *LREC*.

Núria Gala, Thomas François, Delphine Bernhard, and Cédrick Fairon. 2014. Un modèle pour prédire la complexité lexicale et graduer les mots. In *TALN 2014*, pages 91–102.

Núria Gala, Thomas François, and Cédrick Fairon. 2013. Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. *E-lexicography in the 21st century: thinking outside the paper., Tallin, Estonia*.

Martin Gellerstam. 1999. LEXIN-lexikon för invandrare. *LexicoNordica*, (6).

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Bernard Lété, Liliane Sprenger-Charolles, and Pascale Colé. 2004. Manulex: A grade-level lexical database from french elementary school readers. *Behavior Research Methods, Instruments, & Computers*, 36(1):156–166.

Habibollah Mashhady, Behruz Lotfi, and Mahbobeh Noura. 2012. Word Type Effects on L2 Word Retrieval and Learning: Homonym versus Synonym Vocabulary Instruction. *Iranian Journal of Applied Language Studies*, 3(1):97–118.

J Kevin O'Regan and Arthur M Jacobs. 1992. Optimal viewing position effect in word recognition: A challenge to current theory. *Journal of Experimental Psychology: Human Perception and Performance*, 18(1):185.

Robert Östling and Mats Wirén. 2013. Compounding in a Swedish Blog Corpus.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 Task 11: Complex Word Identification. In *SemEval at NAACL-HLT*, pages 560–569.

Matthew Shardlow. 2013. A Comparison of Techniques to Automatically Identify Complex Words. In *ACL (Student Research Workshop)*, pages 103–109.

Edgar A Smith. 1961. Devereux readability index. *The Journal of Educational Research*, 54(8):298–303.

Anaïs Tack, Thomas François, Anne-Laure Ligozat, and Cédrick Fairon. 2016a. Evaluating lexical simplification and vocabulary knowledge for learners of french: Possibilities of using the flelex resource. In *LREC*.

Anaïs Tack, Thomas François, Anne-Laure Ligozat, and Cédrick Fairon. 2016b. Modèles adaptatifs pour prédire automatiquement la compétence lexicale d'un apprenant de français langue étrangère. In *La 23ème Conférence sur le Traitement Automatique des Langues Naturelles (JEP-TALN-RECITAL 2016)*.

Elena Volodina and Sofie Johansson Kokkinakis. 2012. Introducing the Swedish Kelly-list, a new lexical e-resource for Swedish. In *LREC*, pages 1040–1046.

Elena Volodina, Ildikó Pilán, Stian Rødven Eide, and Hannes Heidarsson. 2014. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. In *Proceedings of the third workshop on NLP for computer-assisted language learning at SLTC 2014, Uppsala University*, 107. Linköping University Electronic Press.

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016a. SweLL on the rise: Swedish learner language corpus for european reference level studies. *arXiv preprint arXiv:1604.06583*.

Elena Volodina, Ildikó Pilán, Lorena Llozhi, Baptiste Degryse, and Thomas François. 2016b. SweLLex: second language learners' productive vocabulary. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, Umeå, 16th November 2016*, 130, pages 76–84. Linköping University Electronic Press.

Sanja Štajner, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Anaïs Tack, Seid Muhie Yimam, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, New Orleans, United States. Association for Computational Linguistics.