Coursebook Texts as a Helping Hand for Classifying Linguistic Complexity in Language Learners' Writings

Ildikó Pilán, David Alfter, Elena Volodina
Språkbanken, University of Gothenburg, Sweden
{ildiko.pilan,david.alfter,elena.volodina}@svenska.gu.se

Abstract

We bring together knowledge from two different types of language learning data, texts learners read and texts they write, to improve linguistic complexity classification in the latter. Linguistic complexity in the foreign and second language learning context can be expressed in terms of proficiency levels. We show that incorporating features capturing lexical complexity information from reading passages can boost significantly the machine learning based classification of learner-written texts into proficiency levels. With an F_1 score of .8 our system rivals state-of-the-art results reported for other languages for this task. Finally, we present a freely available web-based tool for proficiency level classification and lexical complexity visualization for both learner writings and reading texts.

1 Introduction

Second or foreign (L2) language learners pass through different development stages commonly referred to as *proficiency levels*. A popular scale of such levels is the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001). As learners advance to higher levels, the complexity of the linguistic input that they are able to comprehend (*receptive* skills) and the output that they produce (*productive* skills) increases in terms of both lexical and grammatical patterns. Although learners' receptive and productive knowledge overlap, they only do so partially, the latter being typically a subset of the former corresponding to a somewhat lower linguistic complexity overall (Barrot, 2015).

In previous work, NLP methods have been successfully applied for assessing separately receptive and productive L2 levels (see section 2). We, on the other hand, hypothesize that, since a shared linguistic content exists between what L2 learners are exposed to (*L2 input texts*, e.g. reading passages from coursebooks) and what they produce (*L2 output texts*, e.g. essays), transferring knowledge from one text type may improve the classification of linguistic complexity levels in the other. We focus on the automatic prediction of CEFR levels for L2 learner essays for a number of reasons. Essay writing is a popular means to assess learners' proficiency level and it is a rather subjective and time-consuming task. Moreover, such data is rather scarce and cumbersome to collect (Volodina et al., 2016). Our target language is Swedish since corpora for both L2 text types are available for this language.

We compare two different strategies aiming at improving L2 essay classification results without additional data of this type: (i) employing a word list based on a coursebook corpus for lexical features, (ii) *domain adaptation* experiments, i.e. training a machine learning model on L2 input texts and using it to classify the essays. We first compare the distribution of words per CEFR levels in the essays using two different word lists and find that a list based on L2 input texts correlates well with the manually assigned CEFR labels of the essays. Using this list in machine learning experiments produces a significant performance boost which exceeds our domain adaptation attempts and compares well also to previously reported results for this task. Finally, we present an online tool for assessing linguistic complexity in L2 Swedish input and output texts that performs a machine learning based CEFR level classification and a lexical complexity analysis supported by a color-enhanced visualization of words per level.

Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity,
pages 120–126, Osaka, Japan, December 11-17 2016.

2 Background

Recently a number of attempts emerged at the classification of CEFR levels in input texts (also known as *L2 readability*) which include, among others, systems for French (François and Fairon, 2012), Portuguese (Branco et al., 2014), Chinese (Sung et al., 2015), Swedish (Pilán et al., 2015), and English (Xia et al., 2016). The same type of classification for learner-written texts remains somewhat less explored. Investigations include Vajjala and Lõo (2014) for Estonian and Hancke (2013) for German reporting an F_1 score of .78 and .71 respectively. The systems above are based on supervised learning methods based on rich feature sets.

Relatively few studies exist in the field of assessing the complexity and quality of L2 texts with the use of domain adaptation. Experiments relying on such methods have been explored so far for transferring essay grading models between writing tasks based on different prompts (Zesch et al., 2015; Phandi et al., 2015), and for L2 readability classification by transferring models trained on texts written for native language users to reading passages aimed at L2 learners (Xia et al., 2016).

3 Receptive and Productive L2 Swedish Corpora

Two corpora with L2 focus are currently available for Swedish: SweLL (Volodina et al., 2016), comprised of L2 output texts in the form of learner essays, and COCTAILL (Volodina et al., 2014) containing L2 coursebooks written by experts for L2 learners. The essays in the **SweLL** corpus were written by adult L2 Swedish learners (with available metadata) and they address a variety of topics. In the case of the coursebook corpus, **COCTAILL**, instead of using it in its entirety, we only include reading passages in our dataset. Other coursebook components whose linguistic annotation may be less reliable (e.g. gapped exercises) are excluded. We derive the CEFR level of the reading texts from the level of the lesson (chapter) they occurr in. It is worth mentioning that these two corpora are *independent* from each other, i.e. the essays written by the learners are not based on, or inspired by, the reading passages.

Both corpora span beginner (A1) to advanced (C1) proficiency with texts manually labeled for CEFR levels, and automatically annotated across different linguistic dimensions. These include lemmatization, part-of-speech (POS) tagging and dependency parsing using the Sparv pipeline[1] (Borin et al., 2012). Since A1 level is rather under-represented in both corpora, we exclude them from our experiments. The distribution of texts per type and CEFR level in our datasets is shown in Table 1, where A2 corresponds to elementary level, B1 to intermediate, B2 to upper intermediate and C1 to advanced level.

Writer (text type)	Unit	A2	B1	B2	C1	Total
Learner (L2 output)	Texts	83	75	74	88	**320**
	Tokens	18,349	29,814	32,691	60,095	**140,949**
Expert (L2 input)	Texts	157	258	288	115	**818**
	Tokens	37,168	79,124	101,297	71,723	**289,312**

Table 1: Overview of CEFR-level annotated Swedish datasets.

4 L2 Lexical Complexity: a Comparison of Word Lists

KELLY (Volodina and Kokkinakis, 2012) is a popular L2 Swedish word lists, compiled based on web corpora. It contains 8,425 headwords with not only frequency information, but also suggested CEFR levels based on normalized frequencies. The list has been successfully applied previously in machine learning experiment for classifying CEFR levels in L2 input texts (Pilán et al., 2015).

SVALex (Francois et al., 2016) is another Swedish word list with an L2 focus, created recently. The list contains word frequencies based on reading passages from COCTAILL (see Section 3), which, however, are not connected to suggested CEFR levels. Therefore, we propose an enhanced version of this list,

[1] `https://spraakbanken.gu.se/sparv/`

SVALex+, that includes mappings from frequency distributions to a single CEFR label following the methodology described in Alfter et al. (2016). To create the mappings, as a first step, frequency counts are normalized. Part of this consists of taking the raw frequency counts from SVALex and calculating *per-million-word* (PMW) frequency distributions for all words. These distributions are complemented with *word diversity* distributions, i.e. information about how often a word is used in different coursebooks at each level in the COCTAILL corpus. The intuition is that, if a word is used often at a certain level, but only in one book, it is less representative of a level than if it appears in several coursebooks. We then combine these two distributions into one single *normalized frequency* ($Freq^n$) value for each word by taking the average of the PMW frequency distribution and the word diversity distribution.

The second step consists of mapping these normalized frequencies to CEFR levels. Rather than mapping to the CEFR level at which a word first appears, we establish a *significant onset of use*, a threshold indicating a difference between normalized frequency distributions that is sufficiently large for a level to qualify as mapping for a word. We set this threshold to 0.4 based on initial empirical investigations with L2 teachers during which the overlap between teacher- and system-assigned levels for a small subset of words have been compared. Thus, we map each word to the lowest CEFR level L for which $Freq^n_L - Freq^n_{L-1} > 0.4$ holds, with $L - 1$ being the previous CEFR level and $Freq^n_{L-1} = 0$ if $L =$ A1.

Table 2 compares the percentage of tokens belonging to different CEFR levels based on KELLY and SVALex+ (rows) per essay CEFR level (columns).

	Essay CEFR levels									
	KELLY					**SVALex+**				
	A1	**A2**	**B1**	**B2**	**C1**	**A1**	**A2**	**B1**	**B2**	**C1**
A1	69.0	72.91	72.56	73.3	70.91	74.28	77.86	65.09	61.96	56.92
A2	4.08	3.96	4.18	4.31	5.22	2.54	3.6	8.01	9.31	10.67
B1	1.2	1.79	1.52	2.4	3.16	1.73	2.91	9.82	12.59	14.28
B2	.67	.68	1.17	.83	1.11	.43	.66	.86	1.07	1.86
C1	.43	.31	.31	.4	.5	.14	.17	.33	.48	.81

Table 2: Distribution of token CEFR levels (in %) per essay CEFR levels.

We can observe that the distribution of tokens per CEFR level based on KELLY remains rather unchanged: A1 and C2 level essays contain, for instance, approximately the same amount of A1-C1 tokens. SVALex+, on the other hand, correlates better with the overall CEFR level of the essays. The highlighted cells show a decrease of lower level tokens in higher level essays and an increase of higher level tokens in more advanced essays. This would suggest that using SVALex+ may improve CEFR level classification performance for learner essays. The amount of B2 and C1 tokens seems still rather limited even at higher levels which can be explained to some extent by SVALex+ containing receptive vocabulary that learners might not be able to use productively.

5 Essay Classification Experiments

5.1 Feature Set

We use the feature set that we described in Pilán et al. (2015) and Pilán et al. (2016) for modeling linguistic complexity in L2 Swedish texts. The 61 features of this set can be divided into five sub-groups: *length-based* (e.g. average sentence and token length), *lexical* (e.g. amount of tokens per CEFR level), *morphological* (e.g. past verbs to verbs ratio), *syntactic* (e.g. average dependency length) and *semantic* features (e.g. number of senses). For a more detailed description of the feature set see the cited works.

5.2 Experimental Setup

We use the sequential minimal optimization algorithm from WEKA (Hall et al., 2009) and the feature set mentioned above for all experiments. Results are obtained using 10-fold cross-validation, unless otherwise specified. Reported measures include F_1 and quadratic weighted kappa (κ^2), a distance-based scoring function taking into consideration also the degree of misclassifications. Our baselines consist of assigning the most frequent label in the dataset to each instance (MAJORITY) and cross-validated results on the learner essays using KELLY (E-KELLY) for lexical features.

Domain adaptation We compare these to two models using information from SVALex+ (E-SVALEX+ with SVALex+ instead of KELLY and E-KELLY&SVALEX+ including both lists), as well as to two simple domain adaptation setups inspired by Daumé III and Marcu (2006). In a domain adaptation scenario, data from a source domain is used to predict labels in a different, target domain. In our SOURCE-ONLY setup, a model trained on coursebook texts is applied to the essays, our target domain. In +FEATURE the CEFR levels predicted by a model trained on coursebook texts is used as an additional feature when training a classifier for the essays. For both the SOURCE-ONLY and the +FEATURE setup the KELLY list has been used.

5.3 Classification Results

The results of our experiments are presented in Table 3.

Essays (baselines)			**Essays (using SVALex+)**			**Coursebooks → Essays**		
	F_1	κ^2		F_1	κ^2		F_1	κ^2
MAJORITY	.120	.000	E-SVALEX+	**.808**	**.922**	SOURCE-ONLY	.438	.713
E-KELLY	.721	.886	E-KELLY&SVALEX+	**.816**	**.930**	+FEATURE	.709	.879

Table 3: Results for different classification improvement strategies.

Substituting KELLY-based features with their SVALex+ based equivalents increases classification performance substantially, from .721 to .822 in terms of F_1. This is most likely connected to the fact that word frequencies based on the general (web) corpus, KELLY, reflect less precisely learners' progression in terms of lexical complexity compared to SVALex+, which is based on texts explicitly intended for L2 learners (see Table 2). Combining both KELLY and SVALex+ achieves a slight gain, but the performance difference remains rather negligible compared to using SVALex+ alone. The high κ^2 values for the SVALex+ based models indicate that very few misclassifications occur with a distance of more than one CEFR level. By inspecting the confusion matrices we find that only two instances fall into this category for the E-SVALex+ model, and none for E-KELLY&SVALex+.

Applying a coursebook model to the essays (SOURCE-ONLY) results in a radical performance drop compared to the in-domain models, which indicates that the distribution of feature values in L2 input and output texts differ to a rather large extent. For the same reason, adding the output of a coursebook based classifier (+FEATURE) performs less accurately than the E-KELLY baseline. These results, however, do not exclude the possibility of a successful transfer between these domains. Additional domain adaptation techniques may be able to bridge the gap between the source and target domain distributions.

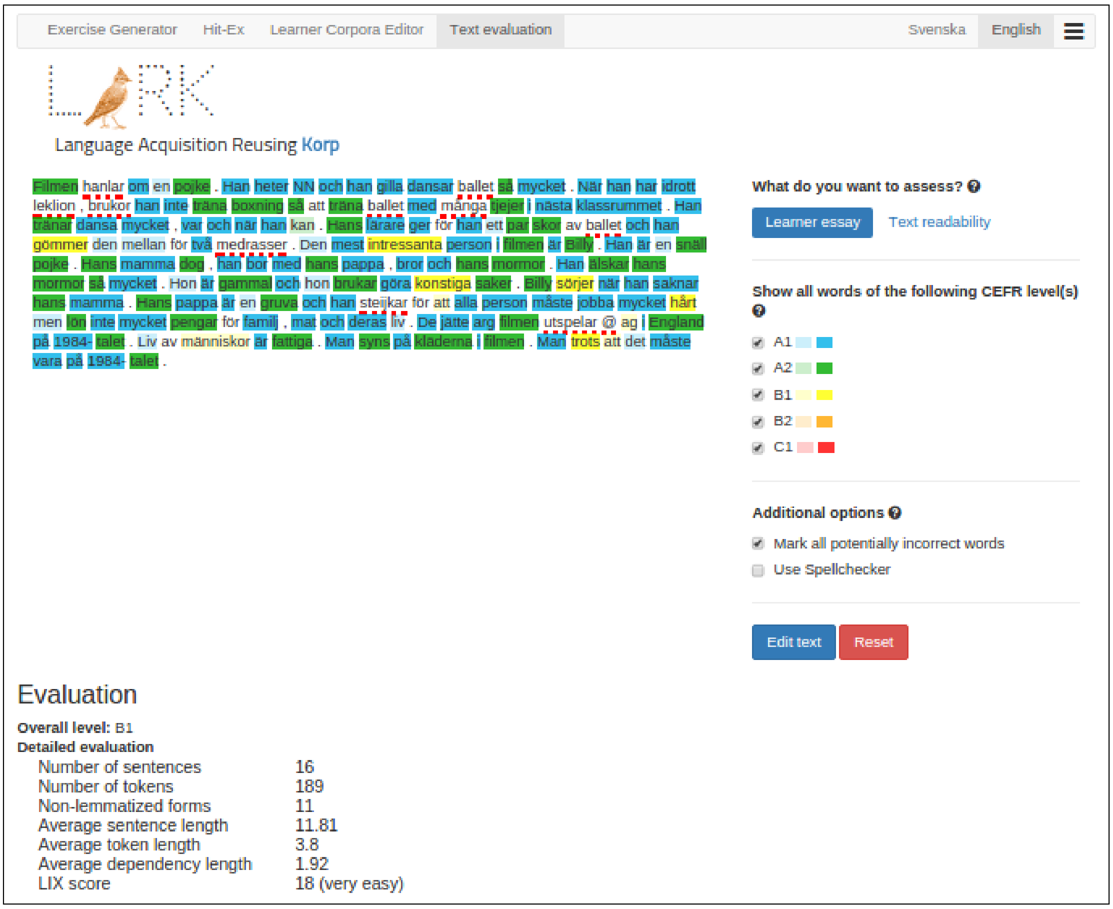
Our SVALex+ based models achieve state-of-the-art performance compared to CEFR level classification systems for other languages such as the German system with .71 F_1 from Hancke (2013), and the Estonian one with .78 F_1 by Vajjala and Lõo (2014). Both of these systems, however, were built using an approximately three times larger annotated in-domain corpus.

6 An Online Tool for L2 Linguistic Complexity Analysis

To put our L2 linguistic complexity analysis methods to practical use, we have made them available as a free online tool[2]. Figure 1 shows the web interface of the current version of our system.

Users can type or paste a text in a text box and indicate whether the text was written by experts as reading material (”Text readability”) or by learners (”Learner essay”). The text is then automatically analyzed in several steps. First, it undergoes an automatic linguistic annotation with Sparv (e.g. POS

[2] `https://spraakbanken.gu.se/larkalabb/texteval`



Filmen hanlar om en pojke . Han heter NN och han gilla dansar ballet så mycket . När han har idrott leklion , brukor han inte träna boxning så att träna ballet med många tjejer i nästa klassrummet . Han tränar dansa mycket , var och när han kan . Hans lärare ger för han ett par skor av ballet och han gömmer den mellan för två medrasser . Den mest intressanta person i filmen är Billy . Han är en snäll pojke . Hans mamma dog , han bor med hans pappa , bror och hans mormor . Han älskar hans mormor så mycket . Hon är gammal och hon brukar göra konstiga saker . Billy sörjer när han saknar hans mamma . Hans pappa är en gruva och han steijkar för att alla person måste jobba mycket hårt men lön inte mycket pengar för familj , mat och deras liv . De jätte arg filmen utspelar @ ag i England på 1984- talet . Liv av människor är fattiga . Man syns på kläderna i filmen . Man trots att det måste vara på 1984- talet .

What do you want to assess?

Learner essay | Text readability

Show all words of the following CEFR level(s)

- ☑ A1
- ☑ A2
- ☑ B1
- ☑ B2
- ☑ C1

Additional options

- ☑ Mark all potentially incorrect words
- ☐ Use Spellchecker

Evaluation

Overall level: B1

Detailed evaluation

Number of sentences	16
Number of tokens	189
Non-lemmatized forms	11
Average sentence length	11.81
Average token length	3.8
Average dependency length	1.92
LIX score	18 (very easy)

Figure 1: The interface for linguistic complexity analysis

tags, dependency relations). Then the annotated text is fed to a machine learning algorithm based on the feature set described in section 5 that assesses the overall linguistic complexity of the text provided in terms of CEFR levels[3]. Some simple statistics and values for traditional readability measures (e.g. average token length, LIX (Björnsson, 1968)) are also included in the final results at the bottom of the page.

In addition to an overall assessment, a detailed visual L2 lexical complexity analysis can be performed. Users can highlight words of different CEFR levels in their text by ticking one (or more) of the check boxes in the right-side menu. The visualization highlights receptive and productive vocabulary items within the same CEFR level with the darker and lighter shade of the same color respectively. The highlighting is based on information from two vocabulary list: SVALex+ for receptive vocabulary and a word list based on SweLL (Alfter et al., 2016), for productive vocabulary, created using the mapping approach described in Section 4 for SVALex+.

7 Conclusions

We described an exploration of different methods to improve the classification of texts produced by L2 Swedish learners into proficiency levels reflecting L2 linguistic complexity. By incorporating information from coursebooks in the form of lexical features indicating the distribution of CEFR levels per token in the texts, we created a system that reaches state-of-the-art performance reported for other languages for this task. Finally, we presented an online tool for linguistic complexity analysis of L2 texts. In the future, additional domain adaptation techniques could be tested for these text types and the effects of incorporating a learner essay based word list on the classification of L2 input texts could also be investigated.

[3] Currently these are in-domain models based on KELLY which we plan to update with the improved (E-SVALex+) model.

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 5th NLP4CALL and 1st NLP4LA workshop, SLTC 2016*, volume No. 130. Linköping Electronic Conference Proceedings.

Jessie Barrot. 2015. Comparing the linguistic complexity in receptive and productive modes. *GEMA Online Journal of Language Studies*, 15(2):65–81.

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

Lars Borin, Markus Forsberg, and Johan Roxendal. 2012. Korp - the corpus infrastructure of Språkbanken. In *LREC*, pages 474–478.

António Branco, João Rodrigues, Francisco Costa, João Silva, and Rui Vaz. 2014. Rolling out text categorization for language learning assessment supported by language technology. *Computational Processing of the Portuguese Language. Springer*, pages 256–261.

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

Hal Daumé III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, pages 101–126.

Thomas Francois, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: a CEFR-graded Lexical Resource for Swedish Foreign and Second Language Learners. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may.

Thomas François and Cédrick Fairon. 2012. An 'AI readability' formula for French as a foreign language. In *Proceedings of the EMNLP and CoNLL 2012*.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: An update. In *The SIGKDD Explorations*, volume 11, pages 10–18.

Julia Hancke. 2013. Automatic prediction of CEFR proficiency levels based on linguistic features of learner language. *Master's thesis, University of Tübingen.*

Peter Phandi, Kian Ming A. Chai, and Hwee Tou Ng. 2015. Flexible Domain Adaptation for Automated Essay Scoring Using Correlated Linear Regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 431–439, Lisbon, Portugal. Association for Computational Linguistics.

Ildikó Pilán, Sowmya Vajjala, and Elena Volodina. 2015. A readable read: Automatic Assessment of Language Learning Materials based on Linguistic Complexity. *In Proceedings of CICLing 2015, to appear in International Journal of Computational Linguistics and Applications*. Available at `http://arxiv.org/abs/1603.08868`.

Ildikó Pilán, Elena Volodina, and Torsten Zesch. 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, Osaka, Japan. Association for Computational Linguistics.

Yao-Ting Sung, Wei-Chun Lin, Scott Benjamin Dyson, Kuo-En Chang, and Yu-Chia Chen. 2015. Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal*, 99(2):371–391.

Sowmya Vajjala and Kaidi Lõo. 2014. Automatic CEFR Level Prediction for Estonian Learner Text. *NEALT Proceedings Series Vol. 22*, pages 113–127.

Elena Volodina and Sofie Johansson Kokkinakis. 2012. Introducing the Swedish Kelly-list, a new lexical e-resource for Swedish. In *Proceedings of 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. *NEALT Proceedings Series Vol. 22*, pages 128–144.

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016. Swell on the rise: Swedish learner language corpus for european reference level studies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).*

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text Readability Assessment for Second Language Learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics.

Torsten Zesch, Michael Wojatzki, and Dirk Scholten-Akoun. 2015. Task-Independent Features for Automated Essay Grading. In *Proceedings of the Building Educational Applications Workshop at NAACL*.