Classification of Swedish learner essays by CEFR levels

Elena Volodina[1], Ildikó Pilán[2], and David Alfter[3]

Abstract. The paper describes initial efforts on creating a system for the automatic assessment of Swedish second language (L2) learner essays from two points of view: holistic evaluation of the reached level according to the Common European Framework of Reference (CEFR), and the lexical analysis of texts for receptive and productive vocabulary per CEFR level. We describe the data and resources that our experiments were based on, provide a short introduction to the algorithm for essay classification and experiment results, present the user interface we developed for testing new essays, and outline future work.

Keywords: automatic online L2 essay classification, lexical complexity assessment, productive and receptive vocabulary by CEFR levels.

1. Introduction

Learner essay grading presents a lot of challenges, especially in terms of manual assessment time and assessors' qualification. Evaluating learner writing quality can be very time-consuming since it stretches along different linguistic dimensions and thus might need several iterations of re-reading. Human assessment is precise and reliable provided that assessors are well trained. However, their judgements may also be subject to different outside factors, such as hunger or a negative attitude to a learner. To avoid misjudgements and to ensure objectivity, certain institutions have started to complement human grading with automatic assessment as a more objective reference point, e.g. Educational Testing Services (Burstein & Chodorow, 2010).

[1] Språkbanken, University of Gothenburg, Gothenburg, Sweden; elena.volodina@svenska.gu.se
[2] Språkbanken, University of Gothenburg, Gothenburg, Sweden; ildiko.pilan@svenska.gu.se
[3] Språkbanken, University of Gothenburg, Gothenburg, Sweden; david.alfter@svenska.gu.se

How to cite this article: Volodina, E., Pilán, I., & Alfter, D. (2016). Classification of Swedish learner essays by CEFR levels. In S. Papadima-Sophocleous, L. Bradley & S. Thouësny (Eds), *CALL communities and culture – short papers from EUROCALL 2016* (pp. 456-461). Research-publishing.net. https://doi.org/10.14705/rpnet.2016.eurocall2016.606

Developing a data-driven Automatic Essay Grading (AEG) system is a non-trivial task which needs to rely on (1) *data* consisting of essays manually graded by human assessors, (2) a *set of rules* or specific *features* relevant for the assessment, and (3) a *classification algorithm* based on the example data provided and the specified features that can predict the grade or level of previously unseen essays. AEG tasks have been addressed previously in a number of projects, e.g. Hancke and Meurers (2013) for German, Burstein and Chodorow (2010) for English, and Vajjala and Lõo (2014) for Estonian. For Swedish, Östling, Smolentzov, Tyrefors Hinnerich, and Höglin (2013) have looked at Swedish upper secondary school essays, i.e. first language (L1) learner essays, and evaluated them in terms of performance grades (pass with distinction, pass, fail). In contrast to them, our main aim has been to assess the reached proficiency levels in essays written by L2 learners of Swedish.

The system presented here uses CEFR levels (Council of Europe, 2001). The CEFR framework has been selected since it is very influential in both Europe and outside with numerous projects targeting its interpretation (e.g. Hancke & Meurers, 2013; Vajjala & Lõo, 2014), however, very little work has been done for CEFR-based L2 Swedish.

2. L2 essay classification

2.1. Essay corpus

The availability of data is critical for AEG experiments. Our experiments are based on SweLL (Volodina et al., 2016), a corpus consisting of L2 Swedish learner essays, linked to proficiency levels as defined by CEFR. Essays cover five of the six proficiency levels (see Table 1) with varying amounts of essays per level.

All essays contain information on learners' mother tongue(s), age, gender, education level, and at which CEFR level the essay is. Essays have been used to extract features based on available annotation, such as level, dictionary forms, word classes, syntactic annotation.

Table 1. Overview of SweLL corpus

	A1	**A2**	**B1**	**B2**	**C1**	**Unknown**	**Total**
Nr essays	16	83	75	74	89	2	339
Nr tokens	2 084	18 349	29 814	32 691	60 095	360	144 087

2.2. Feature selection

Feature selection is the most important and time-consuming part of an AEG project. Features can be language independent, such as n-grams, sentence- and word-length, or language specific, such as out-of-vocabulary words (where vocabulary is defined as some lexicon or word list). Our experiments included an empiric analysis of data, the extraction of relevant features in machine learning experiments and experimentation with those to select the most predictive ones. Our complete set of 61 features (Pilán, Vajjala, & Volodina, forthcoming) extracted from the linguistic annotation available in SweLL include count-based, lexical, syntactic, morphological, and semantic features.

2.3. Essay classification experiments and results

Using SweLL as training data, we created a classification system which predicts which CEFR level the writer of an essay has performed at. We used the Sequential Minimal Optimization (SMO) machine learning algorithm available in WEKA (http://www.cs.waikato.ac.nz/~ml/weka/), which based on the linguistic features observed in hand-annotated essays is able to learn how to automatically assign a CEFR level to a previously unseen essay. Table 2 presents results obtained using different types of features, where F1 is the harmonic mean of precision and recall, and accuracy expresses the amount of correctly classified texts. The number of features per sub-group is also indicated since it may influence performance.

Table 2. Classification results

	Nr features	F1	Accuracy (%)
All	61	0.66	66.96
Count	7	0.45	51.48
Lexical	11	0.58	59.52
Morphological	30	0.53	55.35
Syntactic	11	0.51	53.86
Semantic	2	0.28	36.90

Our system with the complete feature set (ALL) classified essays with 67% accuracy, i.e. making correct assessments about seven out of ten times. However, almost all (98.5%) classification errors were minor, within one CEFR level distance from the teacher-assigned level, a very encouraging result which compares well to the human performance of 45.8% reported in Östling et al. (2013) and systems for other languages using three times more annotated data, e.g. 61% for German (Hancke & Meurers, 2013) and 79% for Estonian (Vajjala & Lõo, 2014). Lexical features were the most informative, and the most useful single features included the number of tokens per CEFR level and word-list based frequency information.

3. Lexical complexity analysis

Both previous research and our experiments have indicated lexical features as one of the most predictive ones (Pilán et al., forthcoming). For this reason, we experimented with a stand-off (i.e. separate from essay classification in section 2) lexical analysis of the essays for giving insights into the lexical complexity of a text, seen from receptive and productive perspectives. Similar efforts have been taken for other languages, e.g. English (http://www.englishprofile.org/wordlists/text-inspector) and French (http://cental.uclouvain.be/flelex/), however, our resources allow us to identify receptive versus productive lexical items per level, whereas only productive vocabulary is targeted in the English system and only receptive in the French one. To be able to perform lexical analysis of texts, two lists have been employed: SVALex (François, Volodina, Pilán, & Tack, 2016) and SweLL-list (Llozhi, 2016).

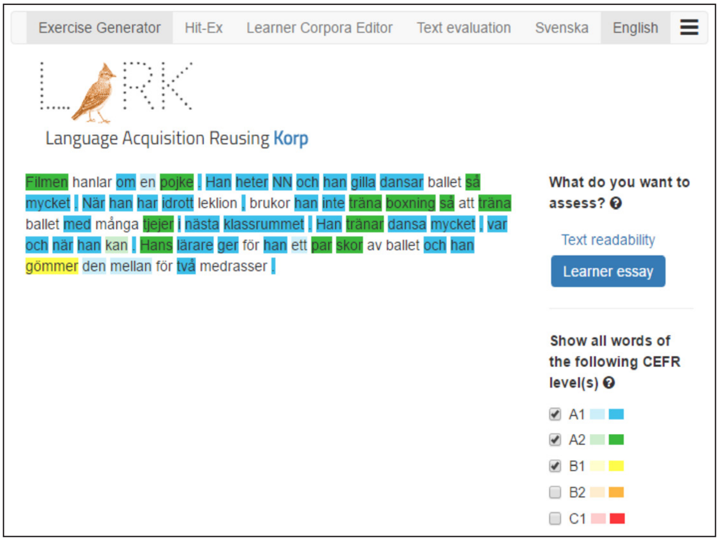
SVALex is a frequency-based list derived from reading comprehension texts used for teaching CEFR courses, thus representing lexical items that L2 learners are exposed to while reading or listening, i.e. receptive vocabulary. The *SweLL list* is derived from the SweLL corpus, showing the distribution of lexical items over CEFR levels based on frequency information. Since SweLL-list items come from essays, they indicate the productive use of vocabulary. Each item in the two lists, a combination of a dictionary form (lemma) and parts of speech, has associated information on levels at which it appears. We preliminarily consider frequency peaks as an indication of the target level for that item. Refinement of the strategies for identifying target levels are under development.

For analysis of *lexical complexity,* each word (i.e. its lemma in combination with its part of speech) in an essay is tested against the two resources and is associated with the CEFR level for receptive or productive knowledge.

4. User interface

The described work has resulted in an online service for testing arbitrary new essays written in Swedish. This is the first prototype of our system, where natural language processing tools are combined to deliver a user-friendly analysis of essays. Initially, essays undergo automatic linguistic analysis which generates dictionary forms, parts of speech and syntactic annotation. Then, depending upon user choices, a holistic assessment (i.e. reached CEFR level) as well as lexical analysis of an essay are generated using resources and techniques described above.

Figure 1. User interface for L2 text classification



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 .

Figure 1 shows available choices on the right, the feedback statement below the input and colour-coded lexical analysis of the pasted texts. Users have a choice to evaluate either L2 learner essays or reading comprehension texts. However, since classification of reading comprehension texts is outside the scope of this paper, we do not go into details. Words from the selected CEFR levels are highlighted showing receptive ones in a lighter color and productive ones in a darker color.

The SweLL-based online system is not yet released for public use, however an experimental version is already available through the Swedish Language Bank at the university of Gothenburg, Sweden (https://spraakbanken.gu.se/larkalabb/).

5. Concluding remarks

Our classification experiments showed that, even though the presented system is an initial prototype and more work needs to be invested to make it fully functional and useful in the language learning context with regards to evaluation of learner-written texts, essay classification results are promising. We found that considering the lexical dimension is particularly effective for CEFR level classification. Further work on refining the SweLL-AEG algorithm would include, among others, adding error annotation to the essays, linking error types to CEFR proficiency levels, and employing error types as a feature in our algorithm. Availability of error annotation would also facilitate a more instructive feedback to learners.

6. Acknowledgements

We thank the Swedish Language Bank (Språkbanken) and Department of Swedish at the University of Gothenburg, Sweden for supporting our work and providing necessary financing.

References

Burstein, J., & Chodorow, M. (2010). Progress and new directions in technology for automated essay evaluation. In R. B. Kaplan (Ed.), *The Oxford handbook of applied linguistic*s (2 ed.). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195384253.013.0036

Council of Europe. (2001). *The common European framework of reference for languages: learning, teaching, assessment*. Cambridge University Press.

François, T., Volodina, E., Pilán, I., & Tack, A. (2016). SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. *Proceedings of LREC 2016, Slovenia*.

Hancke, J., & Meurers, D. (2013). Exploring CEFR classification for German based on rich linguistic modeling. *LCR 2013*.

Llozhi, L. (2016). *SWELL LIST. A list of productive vocabulary generated from second language learners' essays*. Master Thesis. University of Gothenburg.

Pilán, I., Vajjala, S., & Volodina, E. (forthcoming). A readable read: automatic assessment of language learning materials based on linguistic complexity. *To appear in International Journal of Computational Linguistics and Applications (IJLCA)*.

Vajjala, S., Lõo, K. (2014). Automatic CEFR level prediction for Estonian learner text. *Proceedings of the third workshop on NLP for computer-assisted language learning*. NEALT Proceedings Series 22 / Linköping Electronic Conference Proceedings 107.

Volodina, E., Pilán, I., Enström, I., Llozhi, L., Lundkvist, P., Sundberg, G., & Sandell, M. (2016). SweLL on the rise: Swedish learner language corpus for European reference level studies. *Proceedings of LREC 2016, Sloveni*a.

Östling, R., Smolentzov, A., Tyrefors Hinnerich, B., & Höglin, E. (2013). Automated essay scoring for Swedish. *The 8th Workshop on Innovative Use of NLP for Building Educational Applications, US.*

Research-publishing.net

Published by Research-publishing.net, not-for-profit association
Dublin, Ireland; Voillans, France, info@research-publishing.net

CALL communities and culture – short papers from EUROCALL 2016
Edited by Salomi Papadima-Sophocleous, Linda Bradley, and Sylvie Thouësny

Disclaimer: Research-publishing.net does not take any responsibility for the content of the pages written by the authors of this book. The authors have recognised that the work described was not published before, or that it is not under consideration for publication elsewhere. While the information in this book are believed to be true and accurate on the date of its going to press, neither the editorial team, nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, expressed or implied, with respect to the material contained herein. While Research-publishing.net is committed to publishing works of integrity, the words are the authors' alone.

Trademark notice: product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Copyrighted material: every effort has been made by the editorial team to trace copyright holders and to obtain their permission for the use of copyrighted material in this book. In the event of errors or omissions, please notify the publisher of any corrections that will need to be incorporated in future editions of this book.

Typeset by Research-publishing.net
Cover design by © Easy Conferences, info@easyconferences.eu, www.easyconferences.eu
Cover layout by © Raphaël Savina (raphael@savina.net)
Photo "bridge" on cover by © Andriy Markov/Shutterstock
Photo "frog" on cover by © Fany Savina (fany.savina@gmail.com)
Fonts used are licensed under a SIL Open Font License

ISBN13: 978-1-908416-43-8 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

ISBN13: 978-1-908416-44-5 (Ebook, PDF, colour)
ISBN13: 978-1-908416-45-2 (Ebook, EPUB, colour)

Legal deposit, Ireland: The National Library of Ireland, The Library of Trinity College, The Library of the University of Limerick, The Library of Dublin City University, The Library of NUI Cork, The Library of NUI Maynooth, The Library of University College Dublin, The Library of NUI Galway.

Legal deposit, United Kingdom: The British Library.
British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: décembre 2016.