Automatic Selection of Suitable Sentences for Language Learning Exercises

Ildikó Pilán[1], Elena Volodina[2], and Richard Johansson[3]

Abstract. In our study we investigated second and foreign language (L2) sentence readability, an area little explored so far in the case of several languages, including Swedish. The outcome of our research consists of two methods for sentence selection from native language corpora based on Natural Language Processing (NLP) and machine learning (ML) techniques. The two approaches have been made available online within Lärka, an Intelligent CALL (ICALL) platform offering activities for language learners and students of linguistics. Such an automatic selection of suitable sentences can be valuable for L2 teachers during the creation of new teaching materials, for L2 students who look for additional self-study exercises as well as for lexicographers in search of example sentences to illustrate the meaning of a vocabulary item. Members from all these potential user groups evaluated our methods and found the majority of the sentences selected suitable for L2 learning purposes.

Keywords: sentence readability, Swedish, NLP, ICALL, CEFR, GDEX, retrieval, machine learning, supervised classification, corpus-based evidence.

1. Introduction

Native language (L1) texts are a valuable source of authentic sentences suitable for the purposes of L2 learning, either as exercise items or as examples illustrating the meaning of a word. Before being able to use such sentences in CALL systems, however, we have to ensure that these examples are *readable*, i.e. understandable

[1] Språkbanken, University of Gothenburg, Göteborg, Sweden; ildiko.pilan@gmail.com

[2] Språkbanken, University of Gothenburg, Göteborg, Sweden

[3] Språkbanken, University of Gothenburg, Göteborg, Sweden

How to cite this article: Pilán, I., Volodina, E., & Johansson, R. (2013). Automatic Selection of Suitable Sentences for Language Learning Exercises. In L. Bradley & S. Thouësny (Eds.), *20 Years of EUROCALL: Learning from the Past, Looking to the Future. Proceedings of the 2013 EUROCALL Conference, Évora, Portugal* (pp. 218-225). Dublin/Voillans: © Research-publishing.net.

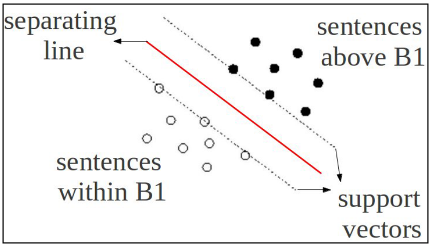
by learners both lexically and structurally. Identifying these sentences manually would require a considerable amount of time. Instead, we propose two automatized selection methods which perform this task for Swedish. Both approaches have been integrated into the online ICALL platform *Lärka* (Volodina, Borin, Loftsson, Arnbjörnsdóttir, & Leifsson, 2012) as part of a sentence readability module called *HitEx* (Hitta Exempel [Find Examples] or Hit Examples). The selection is based on a number of linguistic factors which were found influential for L2 readability, as well as principles of Good Dictionary Examples (GDEX) (Husák, 2008; Kilgarriff, Husák, McAdam, Rundell, & Rychlý, 2008). The sentences selected by the current version of the system have been evaluated by L2 Swedish teachers, learners and linguists, who provided us with positive feedback.

2. Materials and method

The materials used throughout the study included Swedish native language corpora of various genres (novels, newspapers and blog texts) which are accessible through an online tool called *Korp* (Borin, Forsberg, & Roxendal, 2012). Korp offers annotations at different linguistic levels for each sentence including parts of speech (POS), morphosyntactic and syntactic (dependency) relations, which have all been exploited in our selection methods. Furthermore, we employed the scale described in the *Common European Framework of Reference for Languages* (CEFR) when distinguishing L2 difficulty levels. Besides native language corpora, we also utilized the *CEFR corpus* (Volodina, Pijetlovic, Pilán, & Johansson Kokkinakis, 2013), a collection of L2 Swedish materials currently under development, and the *Kelly-list* (Volodina & Johansson Kokkinakis, 2012), a frequency-based word list with CEFR levels for each item. The platform Lärka, besides the HitEx module in which our selection methods have been incorporated, also includes an exercise generator module (Volodina et al., 2013).

The material described above served as basis for our two selection methods: a rule-based and a combined approach using rules as well as ML techniques. As a starting point, we used an algorithm described in Volodina, Johansson, and Johansson Kokkinakis (2012) based on four selection criteria. This initial set of rules was extended with additions from the GDEX literature (Kilgarriff et al. 2008; Husák, 2008), as well as sentence selection research in the L2 context (Segler, 2007) and readability studies for L1 Swedish (Heimann Mühlenbock, 2013; Sjöholm, 2012). The ML method used in the combined approach consisted of *supervised classification*, a process in which our model learned to predict whether a sentence is understandable at B1 (intermediate) proficiency level or not, based on training examples from the CEFR corpus and native language corpora. The classification algorithm employed was a Support Vector Machine (SVM) classifier which aimed at finding a line separating the two classes in the training data (within and above B1 level) based on the linguistic characteristics (*features*) of each sentence. A visual representation of this idea is presented in Figure 1 below.

Figure 1. Support vector machine classification

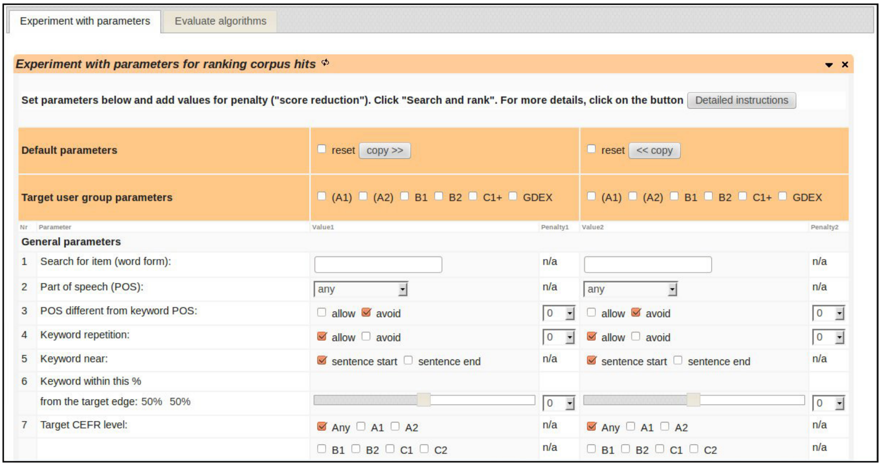


Once trained, the SVM tried to place previously unseen sentences from L1 corpora into the right class. The accuracy of the classifier expresses what percentage of these classifications were correct.

3. HitEx: the L2 sentence readability module

Through the graphical user interface of the HitEx module in Lärka a number of search criteria for the selection of sentences can be set. Figure 2 illustrates part of this page.

Figure 2. The HitEx web page



Experiment with parameters for ranking corpus hits

Set parameters below and add values for penalty ("score reduction"). Click "Search and rank". For more details, click on the button

Nr	Parameter	Value1	Penalty1	Value2	Penalty2
	Default parameters	□ reset		□ reset	
	Target user group parameters	□ (A1) □ (A2) □ B1 □ B2 □ C1+ □ GDEX		□ (A1) □ (A2) □ B1 □ B2 □ C1+ □ GDEX	
	General parameters				
1	Search for item (word form):		n/a		n/a
2	Part of speech (POS):	any	n/a	any	n/a
3	POS different from keyword POS:	□ allow ☑ avoid	0	□ allow ☑ avoid	0
4	Keyword repetition:	☑ allow □ avoid	0	☑ allow □ avoid	0
5	Keyword near:	☑ sentence start □ sentence end	n/a	☑ sentence start □ sentence end	n/a
6	Keyword within this %				
	from the target edge: 50% 50%		0		0
7	Target CEFR level:	☑ Any □ A1 □ A2	n/a	☑ Any □ A1 □ A2	n/a
		□ B1 □ B2 □ C1 □ C2	n/a	□ B1 □ B2 □ C1 □ C2	n/a

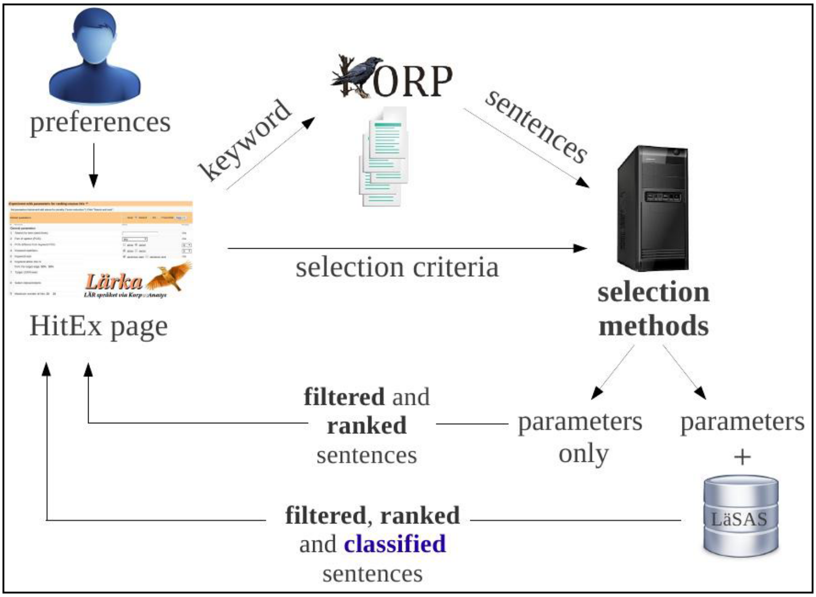
On the left hand side 26 selection criteria, or *parameters*, are listed (only part of these are visible in Figure 2), grouped in three categories: general, structural and lexical. General parameters include basic information about the sentences to select, namely the word to search for (*keyword*), its POS, the corpora from where to choose the examples, etc. Through structural parameters morphosyntactic preferences can be defined. These consist of parameters such as average sentence and word length, percentage of relative pronouns as well as the optional avoidance of participles and modal verbs.

Finally, lexical parameters contain the avoidance of proper names, the allowed percentage of words above the selected CEFR level, etc. Each parameter value can be associated with a *penalty score*, determining the final *ranking* of the sentences based on how well they satisfy the search criteria. A predefined setting is currently available for levels B1, B2, C1+, together with a setting for lexicographers (GDEX). As the presence of the two columns for the parameter values indicates in Figure 2, it is also possible to experiment with two different settings simultaneously.

Instead of using only parameters, the ML component, which we called LäSAS (Lätt/Läs Svenska som Andra Språk [Easy / Read Swedish as a Second Language]), can be selected to be used in combination with some of the parameters. LäSAS classifies sentences based on a large number of linguistic features such as the average number of senses per word, the frequency and CEFR level of words and aspects of syntactic complexity. Such features are based on Swedish L1 readability studies (Heimann Mühlenbock, 2013; Sjöholm, 2012), L2 readability research for other languages (François & Fairon, 2012; Vajjala & Meurers, 2012) and CEFR based course book syllabuses (Levy Scherrer & Lindemalm, 2009). Currently, LäSAS can determine with 70% accuracy whether a sentence is understandable at B1 level or not.

Figure 3 below presents the structure of the readability module and the process of sentence selection. Once users provide their preferences through the dedicated web page in Lärka, the corpus tool, Korp, searches for sentences containing the keyword in Swedish L1 texts. In the next step, sentences undergo a selection with the method previously chosen by the user, which is either purely based on parameters or is a combination of parameters and ML classification with LäSAS. Finally, the resulting filtered set of sentences is displayed on the web page where they can be edited and downloaded to a file. The sentence selection methods are also available as a web service, thus they can be easily integrated in other applications.

Figure 3. The structure of the HitEx readability module



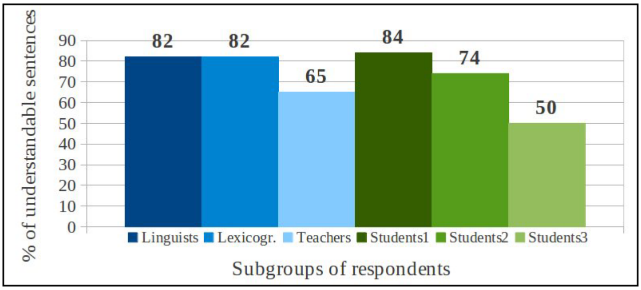
4. Evaluation

To verify whether the sentences selected by our systems are suitable for L2 learning purposes, we carried out an evaluation with altogether 34 participants, including L2 Swedish teachers, students and linguists (including one lexicographer). The respondents had to evaluate a list of 196 sentences chosen with our two selection approaches. Students were required to tell us whether they understood the sentences, whilst teachers and linguists needed to decide whether, according to their judgements, B1 learners would comprehend the sentences. Altogether 73% of the presented items were considered understandable. There was, however, a significant difference among the percentages of understandable examples according to the subgroup of respondents. Figure 4 below shows this discrepancy.

Teachers were considerably stricter than linguists when judging understandability, regarding 17% fewer sentences acceptable. The first subgroup of learners (adults with university-level education) understood 10% more sentences than students above 16 years with mixed educational background (*Students2*) and 34% more than 15-year-old high-school students (*Students3*). Learners understood overall 69% of the examples, 4% more than teachers predicted.

Besides the aspect of understandability, teachers and linguists were also asked to decide whether the sentences would be suitable as exercise items or as examples for vocabulary illustration. About six out of ten sentences corresponded to these criteria. For all three aspects investigated, the purely rule-based approach was slightly preferred (by 3%) to the combined method.

Figure 4. Percentage of understandable sentences per respondent subgroup



During the evaluation, qualitative data has also been collected, which consisted of respondents' comments about difficult or undesirable elements in the sentences. These included, for example, atypical word order, subordinates and the presence of infrequent idioms. Moreover, the lack of sufficient amount of context, informal spelling and a preference for illustrating the most frequent usage of a word have also been mentioned.

5. Conclusions

We proposed two methods for the selection of sentences from native language corpora which are suitable for L2 learning purposes. According to the results of an empirical evaluation, the approach based only on parameters was somewhat more successful than the one combining rules and ML techniques. The results are encouraging, about 70% of the sentences proved to be of an appropriate level of difficulty. About 10% less were suitable as exercise items and example sentences for vocabulary item illustration. The selection methods found their practical application in an ICALL platform in exercise generation and they are also available as a web service. In the future, we intend to extend the selection to all CEFR levels and we also plan to refine the methods further in attempt to improve the suitability of the sentences chosen.

Acknowledgements. We would like to thank all the linguists, teachers and students who completed our evaluation. We also wish to express our appreciation to the publishing house *Liber* for making available materials in electronic format for our study.

References

Borin, L., Forsberg, M., & Roxendal, J. (2012). Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012. Istanbul: ELRA* (pp. 474-478). Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/248_Paper.pdf

François, T., & Fairon, C. (2012). 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* (pp. 466-477). Retrieved from http://aclweb.org/anthology//D/D12/D12-1043.pdf

Heimann Mühlenbock, K. (2013). *I see what you mean – Assessing readability for specific target groups*. PhD Thesis. Data linguistica. University of Gothenburg.

Husák, M. (2008). *Automatic retrieval of good dictionary examples*. Bachelor Thesis. Brno. Retrieved from http://is.muni.cz/th/172590/fi_b/bachelor_thesis.pdf

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proc. Euralex*.

Levy Scherrer, P., & Lindemalm, K. (2009). *Rivstart B1+B2 Textbok*. Stockholm: Natur & Kultur.

Segler, T. M. (2007). *Investigating the selection of example sentences for unknown target words in ICALL reading texts for L2 German*. Doctoral Thesis. University of Edinburgh: Scotland. Retrieved from https://www.era.lib.ed.ac.uk/handle/1842/1750

Sjöholm, J. (2012). *Probability as readability: a new machine learning approach to readability assessment for written Swedish*. Doctoral dissertation. Linköping University, Sweden. Retrieved from http://www.ida.liu.se/projects/webblattlast/Rapporter/lasbarhet.pdf

Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications (BEA7)* (pp. 163-173). Retrieved from http://www.sfs.uni-tuebingen.de/~dm/papers/vajjala-meurers-12.pdf

Volodina, E., Borin, L., Loftsson, H., Arnbjörnsdóttir, B., & Leifsson, G. Ö. (2012). Waste not, want not: towards a system architecture for ICALL based on NLP component re-use. In *Proceedings of the SLTC 2012 workshop on NLP for CALL* (pp. 47-58). Retrieved from http://www.ep.liu.se/ecp/080/006/ecp12080006.pdf

Volodina, E., & Johansson Kokkinakis, S. (2012). Introducing Swedish Kelly-list, a new lexical e-resource for Swedish. In *LREC 2012, Turkey* (pp. 1040-1046). Retrieved from http://www.lrec-conf.org/proceedings/lrec2012/pdf/264_Paper.pdf

Volodina, E., Johansson, R., & Johansson Kokkinakis, S. (2012). Semi-automatic selection of best corpus examples for Swedish: initial algorithm evaluation. In *Proceedings of the SLTC 2012 workshop on NLP for CALL*, Lund. Linköping Electronic Conference Proceedings. Retrieved from http://www.ep.liu.se/ecp/080/007/ecp12080007.pdf

Volodina, E., Pijetlovic, D., Pilán, I., & Johansson Kokkinakis, S. (2013). Towards a gold standard for Swedish CEFR-based ICALL. In *Proceedings of the Second Workshop on NLP for Computer-Assisted Language Learning*. Oslo, Norway. Retrieved from http://www.ep.liu.se/ecp/086/005/ecp13086005.pdf

Research-publishing.net

Published by Research-publishing.net
Dublin, Ireland; Voillans, France
info@research-publishing.net

20 Years of EUROCALL: Learning from the Past, Looking to the Future.
2013 EUROCALL Conference, Évora, Portugal, Proceedings
Edited by Linda Bradley and Sylvie Thouësny

Research-publishing.net has no responsibility for the persistence or accuracy of URLs for external or third-party Internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate. Moreover, 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 (except in the form of an abstract or as part of a published lecture, or thesis), or that it is not under consideration for publication elsewhere. While the advice and information in this book are believed to be true and accurate on the date of its going to press, neither the authors, the editors, 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.

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 editors 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: © Raphaël Savina (raphael@savina.net)
Photos: © Fany Savina (fany.savina@gmail.com)
Fonts used are licensed under a SIL Open Font License

ISBN13: 978-1-908416-12-4 (Paperback, Print on Demand, Lulu.com)
ISBN13: 978-1-908416-13-1 (Ebook, PDF file, Open Access, Research-publishing.net)
ISBN13: 978-1-908416-14-8 (Ebook, Kindle Edition, Amazon Media EU S.à r.l.)
ISBN13: 978-1-908416-15-5 (Ebook, ePUB file, Open Access, Research-publishing.net)

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: novembre 2013.