categories - $<= B1$ and $> B1$, we combined the model's predictions into two classes - A1, A2, B1 were considered as <=B1 and B2, C1 were considered as >B1. The majority baseline for the dataset was 65%, <=B1 being the class with most instances. The model trained on COCTAILL sentences predicted with 73% accuracy teachers' judgments, an 8% improvement over the majority baseline. There was a considerable difference between the precision score of the two classes, which was 85.4% for <=B1, and only 48.5% for >B1.

Previously published results on sentence-level data include [7], who report 66% accuracy for a binary classification task for English and [8] who obtained an accuracy between 78.9% and 83.7% for Italian binary class data using different kinds of datasets. Neither of these studies, however, had a non-native speaker focus. [9] report 71% accuracy for Swedish binary sentence-level classification from an L2 point of view. Both the adjacent accuracy of our sentence-level model (92%) and the accuracy score obtained with that model on SenRead (73%) improve on that score. It is also worth mentioning that the labels in the dataset from [9] were based on the assumption that all sentences in a text belong to the same difficulty level which, being an approximation (as also Figure 1 shows), introduced some noise in that data.

Although more analysis would be needed to refine the sentence-level model, our current results indicate that a rich feature set that considers multiple linguistic dimensions may result in an improved performance. In the future, the dataset could be expanded with more gold-standard sentences, which may improve accuracy. Furthermore, an interesting direction to pursue would be to verify whether providing finer-grained readability judgments is a more challenging task also for human raters.

5 Conclusion and Future Work

We proposed an approach to assess the proficiency (CEFR) level of Swedish L2 course book texts based on a variety of features. Our document-level model, the first for L2 Swedish, achieved an F-score of 0.8, hence, it can reliably distinguish between proficiency levels. Compared to the wide-spread readability measure for Swedish, LIX, we achieved a substantial gain in terms of both accuracy and F-score (46% and 0.6 higher respectively). The accuracy of the sentence-level model remained lower than that of the text-level model, nevertheless, using the complete feature set the system performed 23% and 22% above the majority baseline and LIX respectively. Misclassifications of more than one level did not occur in more than 8% of sentences, thus, in terms of adjacent accuracy, our sentence-level model improved on previous results for L2 Swedish readability [9].

Most notably, we have found that taking into consideration multiple linguistic dimensions when assessing linguistic complexity is especially useful for sentence-level analysis. In our experiments, using only word-frequency features was almost as predictive as a combination of all features for the document level, but the latter made more accurate predictions for sentences, resulting in a 7% difference in accuracy. Besides L2 course book materials, we tested both our document- and sentence-level models also on unseen data with promising results.

In the future, a more detailed investigation is needed to understand the performance drop between document and sentence level. Acquiring more sentence-level annotated data and exploring new features relying on lexical-semantic resources for Swedish would be interesting directions to pursue. Furthermore, we intend to test the utility of this approach in a real-world web application involving language learners and teachers.

References

1. Collins-Thompson, K., Callan, J.P.: A language modeling approach to predicting reading difficulty. In: HLT-NAACL. (2004) 193–200
2. Schwarm, S.E., Ostendorf, M.: Reading level assessment using support vector machines and statistical language models. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics (2005) 523–530
3. Graesser, A.C., McNamara, D.S., Kulikowich, J.M.: Coh-Metrix providing multilevel analyses of text characteristics. Educational Researcher **40** (2011) 223–234
4. Vajjala, S., Meurers, D.: On improving the accuracy of readability classification using insights from second language acquisition. In: Proceedings of the Seventh Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics (2012) 163–173
5. Heimann Mühlenbock, K.: I see what you mean. PhD thesis, University of Gothenburg (2013)
6. Collins-Thompson, K.: Computational assessment of text readability: A survey of current and future research. Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics **6** (2014) 97–135
7. Vajjala, S., Meurers, D.: Assessing the relative reading level of sentence pairs for text simplification. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL-14), Gothenburg, Sweden, Association for Computational Linguistics (2014)
8. Dell'Orletta, F., Wieling, M., Cimino, A., Venturi, G., Montemagni, S.: Assessing the readability of sentences: Which corpora and features? ACL 2014 (2014) 163
9. Pilán, I., Volodina, E., Johansson, R.: Rule-based and machine learning approaches for second language sentence-level readability. In: Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications, Baltimore, Maryland, Association for Computational Linguistics (2014) 174–184
10. Heilman, M.J., Collins-Thompson, K., Callan, J., Eskenazi, M.: Combining lexical and grammatical features to improve readability measures for first and second language texts. In: Proceedings of NAACL HLT. (2007) 460–467
11. Huang, Y.T., Chang, H.P., Sun, Y., Chen, M.C.: A robust estimation scheme of reading difficulty for second language learners. In: Advanced Learning Technologies (ICALT), 2011 11th IEEE International Conference on, IEEE (2011) 58–62
12. Zhang, L., Liu, Z., Ni, J.: Feature-based assessment of text readability. In: Internet Computing for Engineering and Science (ICICSE), 2013 Seventh International Conference on, IEEE (2013) 51–54

13. Salesky, E., Shen, W.: Exploiting morphological, grammatical, and semantic correlates for improved text difficulty assessment. In: Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications, Baltimore, Maryland, Association for Computational Linguistics (2014) 155–162
14. François, T., Fairon, C.: 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, Association for Computational Linguistics (2012) 466–477
15. Branco, A., Rodrigues, J., Costa, F., Silva, J., Vaz, R.: Rolling out text categorization for language learning assessment supported by language technology. In: Computational Processing of the Portuguese Language. Springer (2014) 256–261
16. Velleman, E., van der Geest, T.: Online test tool to determine the CEFR reading comprehension level of text. Procedia Computer Science **27** (2014) 350–358
17. Björnsson, C.H.: Läsbarhet. Liber (1968)
18. Falkenjack, J., Heimann Mühlenbock, K., Jönsson, A.: Features indicating readability in Swedish text. In: Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013). (2013) 27–40
19. Council of Europe: Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge University Press (2001)
20. Volodina, E., Pilán, I., Eide, S.R., Heidarsson, H.: You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a Second Language. NEALT Proceedings Series Vol. 22 (2014) 128
21. Fasth, C., Kannermark, A.: Form i focus: övningsbok i svensk grammatik. Del B. Folkuniv. Förlag, Lund (1997)
22. Volodina, E., Kokkinakis, S.J.: Introducing the Swedish Kelly-list, a new lexical e-resource for Swedish. In: Proceedings of LREC. (2012) 1040–1046
23. Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., Marsi, E.: MaltParser: A language-independent system for data-driven dependency parsing. Natural Language Engineering **13** (2007) 95–135
24. Borin, L., Forsberg, M., Lönngren, L.: SALDO: a touch of yin to WordNet's yang. Language Resources and Evaluation **47** (2013) 1191–1211
25. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. In: The SIGKDD Explorations. Volume 11. (2009) 10–18
26. Volodina, E., Pilán, I., Borin, L., Tiedemann, T.L.: A flexible language learning platform based on language resources and web services. In: Proceedings of LREC 2014, Reykjavik, Iceland (2014)
27. Segler, T.M.: Investigating the selection of example sentences for unknown target words in ICALL reading texts for L2 German. PhD thesis, University of Edinburgh (2007)
28. Pilán, I., Volodina, E., Johansson, R.: Automatic selection of suitable sentences for language learning exercises. In: 20 Years of EUROCALL: Learning from the Past, Looking to the Future. 2013 EUROCALL Conference, 11th to 14th September 2013 Évora, Portugal, Proceedings. (2013) 218–225