A Readable Read: Automatic Assessment of Language Learning Materials based on Linguistic Complexity

Ildikó Pilán[1], Sowmya Vajjala[2], and Elena Volodina[1]

[1] Swedish Language Bank, University of Gothenburg,
40530 Gothenburg, Sweden
{ildiko.pilan,elena.volodina}@svenska.gu.se
[2] LEAD Graduate School, Seminar für Sprachwissenschaft
Universität Tübingen, 72074 Tübingen, Germany
sowmya@sfs.uni-tuebingen.de

Abstract. Corpora and web texts can become a rich language learning resource if we have a means of assessing whether they are linguistically appropriate for learners at a given proficiency level. In this paper, we aim at addressing this issue by presenting the first approach for predicting linguistic complexity for Swedish second language learning material on a 5-point scale. After showing that the traditional Swedish readability measure, Läsbarhetsindex (LIX), is not suitable for this task, we propose a supervised machine learning model, based on a range of linguistic features, that can reliably classify texts according to their difficulty level. Our model obtained an accuracy of 81.3% and an F-score of 0.8, which is comparable to the state of the art in English and is considerably higher than previously reported results for other languages. We further studied the utility of our features with single sentences instead of full texts since sentences are a common linguistic unit in language learning exercises. We trained a separate model on sentence-level data with five classes, which yielded 63.4% accuracy. Although this is lower than the document level performance, we achieved an adjacent accuracy of 92%. Furthermore, we found that using a combination of different features, compared to using lexical features alone, resulted in 7% improvement in classification accuracy at the sentence level, whereas at the document level, lexical features were more dominant. Our models are intended for use in a freely accessible web-based language learning platform for the automatic generation of exercises.

Key words: readability, machine learning, language learning

1 Introduction

Linguistic information provided by Natural Language Processing (NLP) tools has good potential for turning the continuously growing amount of digital text into interactive and personalized language learning material. Our work aims at overcoming one of the fundamental obstacles in this domain of research, namely how to assess the linguistic complexity of texts and sentences from the perspective of second and foreign language (L2) learners.

There are a number of readability models relying on NLP tools to predict the difficulty (readability) level of a text [1–6]. The linguistic features explored so far for this task incorporate information, among others, from part-of-speech (POS) taggers and dependency parsers. Cognitively motivated features have also been proposed, for example, in the Coh-Metrix [3]. Although the majority of previous work focuses primarily on document-level analysis, a finer-grained, sentence-level readability has received increasing interest in recent years [7–9].

The previously mentioned studies target mainly native language (L1) readers including people with low literacy levels or mild cognitive disabilities. Our focus, however, is on building a model for predicting the proficiency level of texts and sentences used in L2 teaching materials. This aspect has been explored for English [10–13], French [14], Portuguese [15] and, without the use of NLP, for Dutch [16].

Readability for the Swedish language has a rather long tradition. One of the most popular, easy-to-compute formulas is LIX (Läsbarthetsindex, 'Readability index') proposed in [17]. This measure combines the average number of words per sentence in the text with the percentage of long words, i.e. tokens consisting of more than six characters. Besides traditional formulas, supervised machine learning approaches have also been tested. Swedish document-level readability with a native speaker focus is described in [5] and [18]. For L2 Swedish, only a binary sentence-level model exists [9], but comprehensive and highly accurate document- and sentence-level models for multiple proficiency levels have not been developed before.

In this paper, we present a machine learning model trained on course books currently in use in L2 Swedish classrooms. Our goal was to predict linguistic complexity of material written by teachers and course book writers for learners, rather than assessing learner-produced texts. We adopted the scale from the Common European Framework of Reference for Languages (CEFR) [19] which contains guidelines for the creation of teaching material and the assessment of L2 proficiency. CEFR proposes six levels of language proficiency: A1 (beginner), A2 (elementary), B1 (intermediate), B2 (upper intermediate), C1 (advanced) and C2 (proficient). Since sentences are a common unit in language exercises, but remain less explored in the readability literature, we also investigate the applicability of our approach to sentences, performing a 5-way classification (levels A1-C1). Our document-level model achieves a state-of-the-art performance (F-score of 0.8), however, there is room for improvement in sentence-level predictions. We plan to make our results available through the online intelligent computer-assisted language learning platform Lärka[3], both as corpus-based exercises for teachers and learners of L2 Swedish and as web-services for researchers and developers.

In the following sections, we first describe our datasets (section 2) and features (section 3), then we present the details and the results of our experiments

[3] http://spraakbanken.gu.se/larka/

in section 4. Finally, section 5 concludes our work and outlines further directions of research within this area.

2 Datasets

Our dataset is a subset of COCTAILL, a corpus of course books covering five CEFR levels (A1-C1) [20]. This corpus consists of twelve books (from four different publishers) whose usability and level have been confirmed by Swedish L2 teachers. The course books have been annotated both content-wise (e.g. exercises, lists) and linguistically (e.g. with POS and dependency tags) [20]. We collected a total of 867 texts (reading passages) from this corpus. We excluded texts that are primarily based on dialogues from the current experiments due to their specific linguistic structure, with the aim of scaling down differences connected to text genres rather than linguistic complexity. We plan to study the readability of dialogues and compare them to non-dialogue texts in the future.

Besides reading passages, i.e. texts, the COCTAILL corpus contains a number of sentences independent from each other, i.e. not forming a coherent text, in the form of lists of sentences and *language examples*. This latter category consists of sentences illustrating the use of specific grammatical patterns or lexical items. Collecting these sentences, we built a sentence-level dataset consisting of 1874 instances. The information encoded in the content-level annotation of COCTAILL (XML tags *list*, *language_example* and the attribute *unit*) enabled us to include only complete sentences and exclude sentences containing gaps and units larger or smaller than a sentence (e.g. texts, phrases, single words etc.). The CEFR level of both sentences and texts has been derived from the CEFR level of the lesson (chapter) they appeared in. In Table 1, columns 2-5 give an overview of the distribution of texts across levels and their mean length in sentences.[4] The distribution of sentences per level is presented in the last two columns of Table 1. COCTAILL contained a somewhat more limited amount of B2 and C1 level sentences in the form of lists and language examples, possibly because learners handle larger linguistic units with more ease at higher proficiency levels.

3 Features

We developed our features based on information both from previous literature [10, 4, 14, 5, 9] and a grammar book for Swedish L2 learners [21]. The set of features can be divided in the following five subgroups: length-based, lexical, morphological, syntactic and semantic features (Table 2).

Length-based (LEN): These features include sentence length in number of tokens (#1) and characters (#4), extra-long words (longer than thirteen characters) and the traditional Swedish readability formula, LIX (see section 1). For the

[4] The number of different books and publishers is reported per each level, some books spanning more levels.