Rule-based and machine learning approaches for second language sentence-level readability

Ildikó Pilán, Elena Volodina and Richard Johansson
Språkbanken, University of Gothenburg
Box 200, Gothenburg, Sweden
{ildiko.pilan, elena.volodina, richard.johansson}@svenska.gu.se

Abstract

We present approaches for the identification of sentences understandable by second language learners of Swedish, which can be used in automatically generated exercises based on corpora. In this work we merged methods and knowledge from machine learning-based readability research, from rule-based studies of Good Dictionary Examples and from second language learning syllabuses. The proposed selection methods have also been implemented as a module in a free web-based language learning platform. Users can use different parameters and linguistic filters to personalize their sentence search with or without a machine learning component assessing readability. The sentences selected have already found practical use as multiple-choice exercise items within the same platform. Out of a number of deep linguistic indicators explored, we found mainly lexical-morphological and semantic features informative for second language sentence-level readability. We obtained a readability classification accuracy result of 71%, which approaches the performance of other models used in similar tasks. Furthermore, during an empirical evaluation with teachers and students, about seven out of ten sentences selected were considered understandable, the rule-based approach slightly outperforming the method incorporating the machine learning model.

1 Introduction and motivation

Despite the fact that there is a vast selection of existing materials, many language teachers opt for completing course syllabuses with either invented examples or authentic resources, customized to the need of specific learners (Howard and Major, 2004). Collections with millions of tokens of digital text are available for several languages today, part of which would offer adequate practice material for learners of a second or foreign language (L2) to develop their skills further. However, a necessary first step representing a major challenge when reusing copora for automatic exercise generation is how to assess the suitability of the available material. In this study, we explored how we could exploit existing Natural Language Processing (NLP) tools and resources for this purpose.

To overcome copyright issues often limiting full-text access to certain corpora, we decided to work with sentences as linguistic unit when assessing the characteristics of suitability and when generating exercise items. Although a large number of studies exist investigating readability, i.e. understandability, at the text level, the sentence level remains little explored. Similarly, the focus of previous investigations has mainly been readability from native language (L1) readers' perspective, but aspects of L2 readability have been less widely studied. To our knowledge no previous research have explored this latter dimension for Swedish before, hence we aim at filling this gap, which can be useful, besides the purposes mentioned above, also in future sentence and text simplification and adaptation tasks.

We propose a rule-based as well as a combination of rule-based and machine learning methods for the identification of sentences understandable by L2 learners and suitable as exercise items. During the selection of linguistic indicators, we have taken into consideration previously studied features of readability (François and Fairon, 2012; Heimann Mühlenbock, 2013; Vajjala and Meurers, 2012), L2 Swedish curricula (Levy Scherrer and Lindemalm, 2009; Folkuniversitet, 2013) and aspects of Good Dictionary Examples (GDEX)

Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications , pages 174–184,
Baltimore, Maryland USA, June 26, 2014.

(Husák, 2010; Kilgarriff et al., 2008), being that we believe they have some properties in common with exercise items. The current version of the machine learning model distinguishes sentences readable by students at an intermediate level of proficiency from sentences of a higher readability level. The approaches have been implemented and integrated into an online Intelligent Computer-Assisted Language Learning (ICALL) platform, Lärka (Volodina et al., 2013). Besides a module where users can experiment with the filtering of corpus hits, a module with inflectional and vocabulary exercises (making use of the selected sentences with our method) is also available. An initial evaluation with students, teachers and linguists indicated that more than 70% of the sentences selected were understandable, and about 60% of them would be suitable as exercise items according to the two latter respondent groups.

2 Background

2.1 Text-level readability

Readability of texts in different languages has been the subject of several studies and they range from simpler formulas, taking into account superficial text properties, to more sophisticated NLP methods. Traditional readability measures for L1 Swedish at the text level include *LIX* (Läsbarthetsindex, "Readability index") (Björnsson, 1968) and the *Nominal Ratio* (Hultman and Westman, 1977). In recent years a number of studies, mostly focusing on the L1 context, appeared which take into consideration linguistic features based on a deeper text processing. Morphosyntactic aspects informative for L1 readability include, among others, parse tree depth, subordination features and dependency link depth (length) (Dell'Orletta et al., 2011). Language models have also been commonly used for readability predictions (Collins-Thompson and Callan, 2004; Schwarm and Ostendorf, 2005). A recently proposed measure, the *Coh-Metrix* (Graesser et al., 2011), aims at a multilevel analysis of texts, inspired by psycholinguistic principles. It measures not only linguistic difficulty, but also cohesion in texts.

Research on L1 readability for Swedish, using machine learning, is described in Heimann Mühlenbock (2013) and Falkenjack et al. (2013). Heimann Mühlenbock (2013) examined readability along five dimensions: surface features, word usage, sentence structure, idea density and human interest. Mean dependency distance, subordinate clauses and modifiers proved good predictors for L1 Swedish.

Although a number of readability formulas exist for native language users, these might not be suitable predictors of L2 difficulty being that the acquisition processes of L1 and L2 present a number of differences (Beinborn et al., 2012). Studies focusing on L2 readability are considerably fewer in the literature. The linguistic features in this context include, among others, relative clauses, passive voice (Heilman et al., 2007) and the number of coordinate phrases per clause (Vajjala and Meurers, 2012). Crossley et al. (2008) applied some Coh-Metrix indicators to English L2 readability. The authors found that lexical coreferentiality, syntactic similarity and word frequency measures outperformed traditional L1 readability formulas. A language-independent approach to L2 readability assessment, using an online machine learning algorithm, is presented by Shen et al. (2013) which, however, employed only the surface features of average sentence and word length, and word frequencies as lexical feature. The authors found that none of the features in isolation was able to clearly distinguish between the levels.

In the second language teaching scenario, a widely used scale is the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001), which, however, has been less frequently adopted so far in readability studies. The CEFR guidelines for L2 teaching and assessment define six different proficiency levels: A1 (beginner), A2 (elementary), B1 (intermediate), B2 (upper intermediate), C1 (advanced) and C2 (proficiency). François and Fairon (2012) proposed a CEFR-based readability formula for L2 French. Some of the predictive features proved to be structural properties, including shallow length features as well as different morpho-syntactic categories (e.g. present participles) and the presence of words in a list of easy words.

2.2 Sentence-level readability

Many of the text readability measures mentioned above have shortcomings when used on very short passages containing 100 words or less (Kilgarriff et al., 2008). The concept of readability at the sentence level can be related to the selection of appropriate vocabulary example sentences. GDEX

(Husák, 2010; Kilgarriff et al., 2008) is a sentence evaluation algorithm, which, on the basis of lexical and syntactical criteria, automatically ranks example candidates from corpora. Some of the influential linguistic aspects of appropriate example sentences are: their length and structure, the presence of short and common vocabulary items which do not need disambiguation and the absence of anaphoric pronouns. Segler (2007) focuses on the L2 rather than on the lexicographic context. He explores the characteristics of helpful vocabulary examples to be used via an ICALL system for L2 German and underlines the importance of syntactic complexity. Research about ranking Swedish corpus examples is presented in Volodina et al. (2012b). Their first algorithm includes four heuristic rules concerning sentence length, infrequent lexical items, keyword position and the presence of finite verbs, complemented by a sentence similarity measure in the second algorithm. Readability experiments focusing at the sentence level have started to appear recently both for language learning purposes (Pilán et al., 2013) and for detecting differences between simplified and unsimplified sentence pairs (Vajjala and Meurers, 2014).

3 Resources

Our sentence selection module utilizes a number of tools, resources and web services available for Swedish. *Korp*[1], an infrastructure for accessing and maintaining corpora (Borin et al., 2012), contains a large number of Swedish texts which are equipped with automatic annotations (with some exceptions) for part-of-speech (POS), syntactic (dependency) relations, lemma forms and sense ids. Korp offers, among others, a web service for concordances, which makes a search in corpora based on a query (e.g. a keyword and its POS) and returns hits with a sentence-long context. Moreover, with the corpus pipeline of Korp, tools for automatically annotating corpora are also available. A variety of different modern Swedish corpora from Korp have been used throughout this study including novel, newspaper and blog texts.

Another source for sentences was the *CEFR corpus* (Volodina and Johansson Kokkinakis, 2013), a collection of CEFR-related L2 Swedish course book texts. The corpus contains: (a) manual annotations indicating the structure of each lesson in the book (exercises, instructions, texts etc.); (b) automatic linguistic annotations obtained with the annotation tools available through Korp. The CEFR corpus at the time of writing included B1 texts from three course books and B2 texts from one course book. The annotation of additional material covering other CEFR levels was ongoing.

Not only corpora, but also information from frequency word lists has been used for determining the appropriateness of a sentence. The *Kelly list* (Volodina and Kokkinakis, 2012) is a frequency-based vocabulary list mostly built on a corpus of web texts from 2010. Besides frequency information, an associated CEFR level is available for each item. Another frequency-based word list employed for the machine learning experiments is the *Wikipedia list* (Volodina et al., 2012b). It contains the POS and the number of occurrences for each word form in a corpus of Swedish Wikipedia texts.

A central resource of the present study is *Lärka*[2] (Volodina et al., 2013), a freely available online ICALL platform. Currently its exercise generator module offers tasks both for students of linguistics and learners of L2 Swedish (Figure 1). Additional parts include a corpus editor used for the annotation of the CEFR corpus and the sentence selection module presented in this paper, *Hit-Ex*[3] (*Hitta Exempel*, "Find Examples" or Hit Examples). The version under development contains also dictation and spelling exercises (Volodina et al., 2013).

4 Machine learning experiments for readability

4.1 Dataset

We distinguished two different classes in the dataset for the machine learning experiments: (a) sentences understandable at (*within*) B1 level and (b) sentences *above* B1 level. For the former group, sentences were collected from B1-level texts from the CEFR corpus. Sentences above B1 level consisted partly of B2-level sentences from the CEFR corpus, and partly of native language sentences from Korp retrieved on the basis of keywords between B2 and C2 levels according to the Kelly list. Only sentences between the length of 5 and 30 tokens were collected from all resources to decrease the influence of sentence length on the decisions made by the classifiers and to increase the importance of other linguistic features. The

[1] http://spraakbanken.gu.se/korp/

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

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

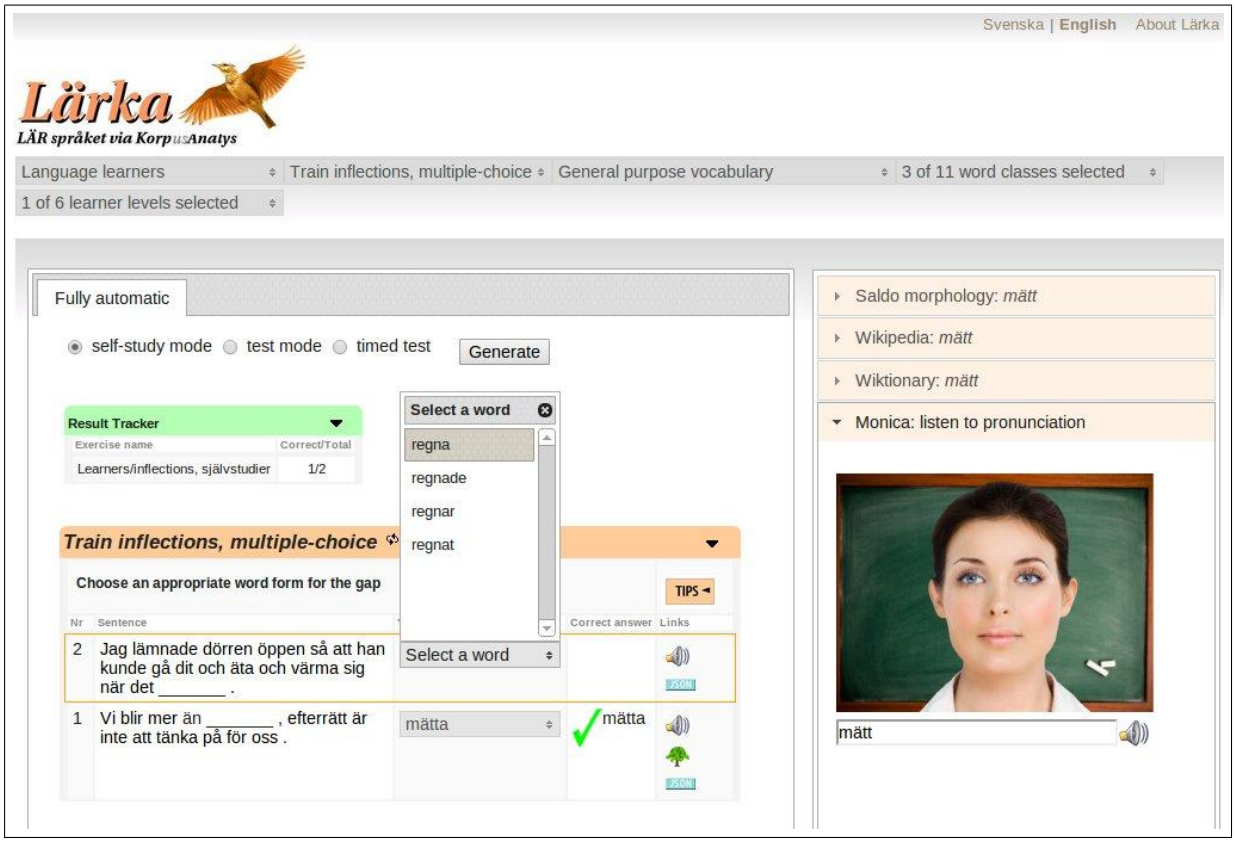

Figure 1: Inflectional exercise.

size of the dataset and the number of sentences per level are illustrated in Table 1.

Level	Source	Nr. sentences
Within B1	B1 (CEFR) texts	2358
Above B1	B2 (CEFR) texts	795
	Korp corpora	1528
Total size of dataset		**4681**

Table 1: The source and the number of sentences in the dataset.

4.2 Method

We performed supervised classification using as training and test data the set of sentences described in section 4.1. Thus, we aimed at a two-way classification distinguishing sentences within B1 level from those above. This level, besides being approximately a middle point of the CEFR scale, is typically divided into sub-levels in language courses (Folkuniversitet, 2013) which indicates a more substantial linguistic content. Consequently, additional practice for learners can be beneficial at this stage. Self-study activities may also be more common in this phase since students have sufficient L2 autonomy. We experimented with different classification algorithms[4] available through the Scikit-learn Python package (Pedregosa et al., 2011), out of which we present the results only of the best performing one here, a linear Support Vector Machine (SVM) classifier. The SVM classifier aims at separating instances into classes with a hyperplane (Tanwani et al., 2009), equivalent to a line in a two-dimensional space. This hyperplane is defined based on the feature values of instances and weights associated with them. Once extracted, the values for each feature were scaled and centered.

Evaluation was carried out with stratified 10-fold cross-validation, i.e. the proportion of labels in each fold was kept the same as that in the whole training set during the ten iterations of training and testing. The evaluation measures taken into consideration were accuracy, precision, recall and the F1 score, a combination of precision and recall, the two of them being equally important (Pedregosa et al., 2011).

[4]The other classification methods used were a Naïve Bayes classifier, a decision tree and two linear algorithms: perceptron and logistic regression.

4.3 Features

After a thorough overview of the machine learning approaches for readability in the literature, a number of features were chosen to be tested in our experiments. The features selected aimed at a deep analysis of the sentences at different linguistic levels. Besides traditional readability indicators, a number of syntactic, morphological, lexical and semantic aspects have been taken into consideration. Our initial set contained altogether 28 features, as presented in Table 2 on the next page.

A number of popular traditional (shallow) features were included in the feature set (features 1-4). These required less sophisticated text processing and had previously been used in several studies with success (Beinborn et al., 2012; Dell'Orletta et al., 2011; François and Fairon, 2012; Heimann Mühlenbock, 2013; Vajjala and Meurers, 2012). We computed sentence length as the number of tokens including punctuation, and token length as the number of characters per token.

Part of the syntactic features was based on the depth (length) and direction of dependency arcs (features 5-8). Another group of these features relied on the type of dependency relations. In feature 9 (Mod) nominal pre-modifiers (e.g. adjectives) and post-modifiers (e.g. relative clauses, prepositional phrases) were counted, similarly to Heimann Mühlenbock (2013). Variation features (ModVar, AdvVar) measured the ratio of a morphosyntactic category to the number of lexical (content) words in the sentence, as in Vajjala and Meurers (2012). These lexical categories comprised nouns, verbs, adverbs and adjectives. Subordinates (11) were detected on the basis of the "UA" (subordinate clause minus subordinating conjunction) dependency relation tag (Heimann Mühlenbock, 2013). Features DepDepth, Mod, Sub and RightDep, PrepComp have previously been empoyed for Swedish L1 readability at the text level in Heimann Mühlenbock (2013) and Falkenjack et al. (2013) respectively.

The lexical-morphological features (features 13-25) constituted the largest group. Difficulty at the lexical level was determined based on both the TTR feature mentioned above, expressing vocabulary diversity, and on the basis of the rarity of words (features 13-17) according to the Kelly list and the Wikipedia word list. An analogous approach was adopted also by François and Fairon (2012), Vajjala and Meurers (2012) and Heimann Mühlenbock (2013) with positive results. The LexD feature considers the ratio of lexical words (nouns, verbs, adjectives and adverbs) to the sum of tokens in the sentence (Vajjala and Meurers, 2012). The NN/VB ratio feature, which has a higher value in written text, can also indicate a more complex sentence (Biber et al., 2004; Heimann Mühlenbock, 2013). Features 21-25 are based on evidence from the content of L2 Swedish course syllabuses (Folkuniversitet, 2013) and course books (Levy Scherrer and Lindemalm, 2009), part of them being language-dependent, namely S-VB/VB and S-VB%. These two features cover different types of Swedish verbs ending in *-s* which can indicate either a reciprocal verb, a passive construction or a deponent verb, active in meaning but passive in form (Fasth and Kannermark, 1989).

Our feature set included three semantic features (26-28). The intuition behind 28 is that words with multiple senses (polysemous words), increase reading complexity as, in order to understand the sentence, word senses need to be disambiguated (Graesser et al., 2011). This feature was computed by counting the number of sense IDs per token according to a lexical-semantic resource for Swedish, SALDO (Borin et al., 2013), and dividing this value by the number of tokens in the sentence. As pronouns indicate a potentially more difficult text (Graesser et al., 2011), we included PN/NN in our set. Both NomR and PN/NN capture idea density, i.e. how complex the relation between the ideas expressed are (Heimann Mühlenbock, 2013).

4.4 Classification results

The results obtained using the complete set of 28 features is shown in Table 3. The results of the SVM are presented in comparison to a baseline classifier assigning the most frequent output label in the dataset to each instance.

Classifier	Acc	F1	B1 Prec	B1 Recall
Baseline	0.50	0.66	0.50	1.00
SVM	0.71	0.70	0.73	0.68

Table 3: Classification results with the complete feature set.

The baseline classifier tagged sentences with 50% accuracy being that the split between the two

Nr.	Feature Name	Feature ID
	Traditional	
1	Sentence length	SentLen
2	Average token length	TokLen
3	Percentage of words longer than 6 characters	LongW%
4	Type-token ratio	TTR
	Syntactic	
5	Average dependency depth	DepDepth
6	Dependency arcs deeper than 4	DeepDep
7	Deepest dependency / sentence length	DDep / SentLen
8	Ratio of right dependency arcs	RightDep
9	Modifiers	Mod
10	Modifier variation	ModVar
11	Subordinates	Sub
12	Prepositional complements	PrepComp

Nr.	Feature Name	Feature ID
	Lexical-morphological	
13	Average word frequency (Wikipedia list)	WikiFr
14	Average word frequency (Kelly list)	KellyFr
15	Percentage of words above B1 level	DiffW%
16	Number of words above B1 level	DiffWs
17	Percentage of words at B1 level	B1W%
18	Lexical density	LexD
19	Nouns/verbs	NN/VB
20	Adverb variation	AdvVar
21	Modal verbs / verbs	MVB/VB
22	Participles / verbs	PCVB/VB
23	S-verbs / verbs	S-VB/VB
24	Percentage of S-verbs	S-VB%
25	Relative pronouns	RelPN
	Semantic	
26	Nominal ratio	NomR
27	Pronoun/noun	PN/NN
28	Average number of senses per word	Sense/W

Table 2: The complete feature set.

classes was about 50-50%. The SVM classified 7 out of 10 sentences accurately. The precision and recall values for the identification of B1 sentences was 73% and 68%. Previous classification results for a similar task obtained an average of 77.25% of precision for the classification of easy-to-read texts within an L1 Swedish text-level readability study (Heimann Mühlenbock, 2013). Another classification at the sentence level, but for Italian and from an L1 perspective achieved an accuracy of 78.2%, thus 7% higher compared to our results (Dell'Orletta et al., 2011). The 73% precision of our SVM model for classifying B1 sentences was close to the precision of 75.1% obtained for the easy-to-read sentences from Dell'Orletta et al. (2011). François and Fairon (2012) in a classification study from the L2 perspective, aiming at distinguishing all 6 CEFR levels for French at the text level, concluded that intermediate levels are harder to distinguish than the levels at the edges of the CEFR scale. The authors reported an adjacent accuracy of 67% for B1 level, i.e. the level of almost 7 out of 10 texts was predicted either correctly or with only one level of difference compared to the original level. Precise comparison with previous results is, however, difficult since, to our knowledge, there are no results reported for L2 readability at the sentence level. Thus, the values mentioned above serve more as a side-by-side illustration.

Besides experimenting with the complete feature set, groups of features were also separately tested. The results are presented in Table 4.

Feature group (Nr of features)	Acc	F1
Traditional (4)	0.59	0.55
Syntactic (8)	0.59	0.54
Lexical (13)	**0.70**	**0.70**
Semantic (3)	0.61	0.55

Table 4: SVM results per feature group.

The group of traditional and syntactic features performed similarly, with an accuracy of 59%. In-

Rank	Feature ID	Weight
1	DiffW%	0.576
2	Sense/W	0.438
3	DiffWs	0.422
4	SentLen	0.258
5	Mod	0.223
6	KellyFr	0.215
7	NomR	0.132
8	AdvVar	0.114
9	Ddep/SentLen	0.08
10	DeepDep	0.08

Table 5: The 10 most informative features according to the SVM weights.

terestingly, although semantic features represented the smallest group, they performed 2% better than traditional or syntactic features. The largest group of features including lexical-morphological indicators performed around 10% more accurately than other feature groups.

Among the 10 features that influenced most the decisions of our SVM classifier, we can find attributes from different feature groups. The ID of these features together with the SVM weights are reported in Table 5. An informative traditional measure was sentence length, similarly to the results of previous studies (Beinborn et al., 2012; Dell'Orletta et al., 2011; François and Fairon, 2012; Heimann Mühlenbock, 2013; Vajjala and Meurers, 2012). Lexical-morphological features based on information about the frequency and the CEFR level of items in the Kelly list (DiffW%, DiffWs and KellyFr) also proved to be influential for the classification, as well as AdvVar. Two out of our three semantic features, namely NomR and, in particular, Sense/W, were also highly predictive. Syntactic features Ddep/SentLen and DeepDep, based on information about dependency arcs, were also among the ten features with highest weights, but they were somewhat less useful, as the weights in Table 5 show.

Contrary to our results, François and Fairon (2012) found syntactic features more informative than semantic ones for L2 French. This may depend either on the difference between the features used or the target languages. Moreover, in the case of Swedish L1 text readability the noun/pronoun ratio and modifiers proved to be indicative of text-level difficulty (Heimann Mühlenbock, 2013), but at the sentence level from the L2 perspective only the latter seemed influential in our experiments.

The data used for the experiments was labeled for CEFR levels at the text level, not at the sentence level. This introduced some noise in the data and made the classification task somewhat harder. In the future, the availability of data labeled at the sentence level could contribute to more accurate results. Excluding potentially lower level sentences from those appearing in higher level texts based on the distance between feature vectors could also be explored, in a similar fashion to Dell'Orletta et al. (2011).

5 Heuristics: GDEX parameters for sentence filtering and ranking

Besides SVM classification, our sentence selection module, Hit-Ex, offers also a number of heuristic parameter options[5], usable either in combination or as an alternative to the machine learning model (for further details see section 6). Part of these search parameters are generic preferences including the keyword to search for, its POS, the corpora from Korp to be used during selection and the desired CEFR level of the sentences. Furthermore, it is possible to avoid sentences containing: abbreviations, proper names, keyword repetition, negative formulations (*inte* "not" or *utom* "except" in the sentence), modal verbs, participles, s-verbs and sentences lacking finite verbs. Users can also allow these categories and choose a penalty point between 0 and -50 for them. The penalty score for each filtering criteria is summed for obtaining a final score per sentence, based on which a final ranking is produced for all sentences retrieved from Korp, the ranking reflecting the extent to which they satisfy the search criteria. Some additional parameters, partly overlapping with the machine learning model's features, are also available for users to experiment with, being that the machine learning model does not cover all CEFR levels. Based on statistical evidence from corpora, we suggested default values for all parameters for retrieving sentences of B1, B2, C1 level with rule-based parameters only. However, additional data and further testing is required to verify the appropriateness of the proposed values.

[5] See Pilán (2013) or the Hit-Ex webpage, http://spraakbanken.gu.se/larka/larka_hitex_index.html, for a complete list of parameters.

6 Combined approach

As mentioned in the previous subsection, the heuristic parameters and the machine learning approach have been implemented and tested also in combination. Parameters are kept to perform a GDEX-like filtering, whilst the SVM model is employed to ensure that hits were of a suitable level for learners. During this combined filtering, first a ranking for each unfiltered sentence coming from the web service of Korp is computed with heuristics. During these calculations, the parameters partly or fully overlapping with certain features of the machine learning model are deactivated, i.e. receive penalty points set to 0, thus, they do not influence the ranking. Instead, those aspects are taken care of by the machine learning model, in a subsequent step. Only the 100 sentences ranked highest are given for classification to the machine learning model for efficiency reasons. Finally, once the classification has been performed, sentences classified as understandable at B1 level are returned in the order of their heuristic ranking. Figure 2 shows part of the interface of Hit-Ex, as well as the highest ranked three sentences[6] of an example search for the noun *hund* "dog" at B1 level. Besides the Hit-Ex page, both the heuristics-only and the combined approaches are available also as web services.

7 Evaluation

The purpose of the evaluation was to explore how many sentences, collected from native language corpora in Korp with our algorithms, were understandable at B1 level (at B1 or below) and thus, appropriate to be presented to learners of L2 Swedish of that CEFR level. Participants included three L2 Swedish teachers, twenty-six L2 Swedish students at B1 level, according to their current or most recent language course, and five linguists familiar with the CEFR scale. Besides the criteria of understandability (readability), the aspect of being an appropriate exercise item was also explored. We selected altogether 196 sentences using both our approaches, with two different parameter settings for the rule-based method (See Pilán et al. (2013) and Pilán (2013) for further details about the evaluation). Evaluators were asked to indicate whether they found the sentences understandable at B1 level or not. Teachers and linguists (TL) rated the sentences also as potential exercise items. The results of the evaluation are presented in Table 6.

Understandability		**Exercise item**
TL	Students	TL
76%	69%	**59%**
73%		

Table 6: Evaluation results.

Respondents found overall 73% percent of the sentences selected by both our methods understandable at B1 level, whilst somewhat less, about six out of ten items, proved to be suitable for being included in exercises for L2 Swedish learning.

According to our evaluators, the two settings of the rule-based approach (Alg1-s1 and Alg1-s2) satisfied the two criteria observed between 1-5% more of the cases. On average, teachers, linguists and students considered 75% of the sentences selected with Alg1-s1 understandable, but only 70% of those identified with the combined approach (Alg2). The detailed results per algorithm, criteria and user group are shown in Figure 3.

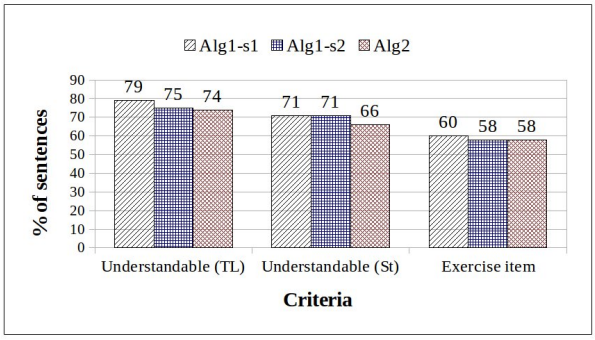

Figure 3: Comparison of algorithms.

According to our evaluators' comments, some of the selected sentences contained difficult aspects at the syntactic level, among others, difficult word order, subordinates and relative clauses. Moreover, at the lexical level, a stricter lexical filtering, and checking for a sufficient amount of lexical words in the sentence would be required. Respondents' comments revealed also the potential future improvement of filtering for context dependency which would make sentences more suitable as exercise items.

[6]English translations of the selected sentences: (1)"It would be enough for a normal dog."; (2)"They left the body in the form of a dog."; (3)"There was a person with a dog."

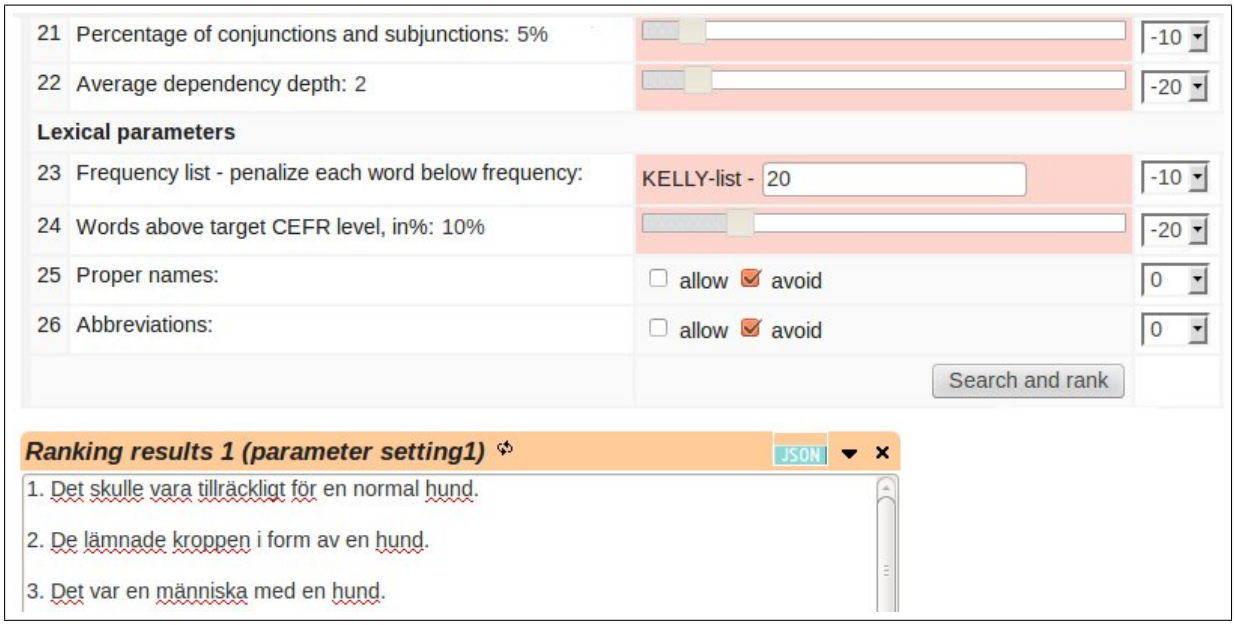

Figure 2: Part of the user interface and example search results.

8 Conclusion

In this study we investigated linguistic factors influencing the sentence-level readability of Swedish from a L2 learning point of view. The main contribution of our work consists of two sentence selection methods and their implementation for identifying sentences from a variety of Swedish corpora which are not only readable, but potentially suitable also as automatically generated exercise items for learners at intermediate (CEFR B1) level and above. We proposed a heuristics-only and a combined selection approach, the latter merging rule-based parameters (targeting mainly the filtering of "undesired" linguistic elements), and machine learning methods for classifying the readability of sentences from L2 learners' perspective. We obtained a classification accuracy of 71% with an SVM classifier which compares well to previously reported results for similar tasks. Our results indicate the success of lexical-morphological and semantic factors over syntactic ones in the L2 context. The most predictive indicators include, besides sentence length, the amount of difficult words in the sentence, adverb variation, nominal pre- and post-modifiers and two semantic criteria, the average number of senses per word and nominal ratio (Table 5). Within a smaller-scale evaluation, about 73% of the sentences selected by our methods were understandable at B1 level, whilst about 60% of the sentences proved to be suitable as exercise items, the heuristics-only approach being slightly preferred by evaluators. Further investigation of the salient properties of exercise items may contribute to the improvement of the current selection approach. The method, as well as most of the parameters and features used, are language independent and could, thus, be applied also to languages other than Swedish, provided that NLP tools performing similarly deep linguistic processing are available. Future additions to the filtering parameters may include aspects of word order, independence from a wider context, valency information and collocations. The optimization of the classifier could also be studied further; different algorithms and additional features could be tested to improve the classification results. The machine learning approach might show improvements in the future with training instances tagged at the sentence level and it can be easily extended, once additional data for other CEFR levels becomes available. Finally, additional evaluations could be carried out to confirm the appropriateness of the sentences ranked by the extended and improved selection method. To indicate the extent to which a sentence is understandable, 4- or 5-point scales may be used, and the employment of exercises instead of a list of sentences to read could also be investigated for verifying the suitability of the examples.

References

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2012. Towards fine-grained readability measures for self-directed language learning. In *Electronic Conference Proceedings*, volume 80, pages 11–19.

Douglas Biber, Susan Conrad, Randi Reppen, Pat Byrd, Marie Helt, Victoria Clark, Viviana Cortes, Eniko Csomay, and Alfredo Urzua. 2004. *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus*. Test of English as a Foreign Language.

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

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

Lars Borin, Markus Forsberg, and Lennart Lönngren. 2013. SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Kevyn Collins-Thompson and James P Callan. 2004. A language modeling approach to predicting reading difficulty. In *HLT-NAACL*, pages 193–200.

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

Scott A Crossley, Jerry Greenfield, and Danielle S McNamara. 2008. Assessing text readability using cognitively based indices. *Tesol Quarterly*, 42(3):475–493.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read-it: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83. Association for Computational Linguistics.

Johan Falkenjack, Katarina Heimann Mühlenbock, and Arne Jönsson. 2013. Features indicating readability in Swedish text. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 27–40.

Cecilia Fasth and Anita Kannermark. 1989. *Goda grunder*. Folkuniversitetets Förlag.

Folkuniversitet. 2013. Kurser i svenska. Svenska B1. `http://www.folkuniversitetet.se/Kurser--Utbildningar/Sprakkurser/Svenska_Swedish/Svenska-B1--Swedish-B1/`.

Thomas François and Cédrick Fairon. 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*, pages 466–477. Association for Computational Linguistics.

Arthur C Graesser, Danielle S McNamara, and Jonna M Kulikowich. 2011. Coh-Metrix providing multilevel analyses of text characteristics. *Educational Researcher*, 40(5):223–234.

Michal J. Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of NAACL HLT*, pages 460–467.

Katarina Heimann Mühlenbock. 2013. *I see what you mean*. Ph.D. thesis, University of Gothenburg.

Jocelyn Howard and Jae Major. 2004. Guidelines for designing effective English language teaching materials. In *9th Conference of Pan Pacific Association of Applied Linguistics*.

Tor G Hultman and Margareta Westman. 1977. *Gymnasistsvenska*. Liber.

Milos Husák. 2010. *Automatic retrieval of good dictionary examples*. Bachelor Thesis, Brno.

Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlỳ. 2008. GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of Euralex*.

Paula Levy Scherrer and Karl Lindemalm. 2009. *Rivstart B1 + B2. Textbok*. Natur och Kultur, Stockholm.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, 12:2825–2830.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2013. 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.*, pages 218–225.

Ildikó Pilán. 2013. *NLP-based Approaches to Sentence Readability for Second Language Learning Purposes*. Master's Thesis, University of Gothenburg. `https://www.academia.edu/6845845/NLP-based_Approaches_to_Sentence_Readability_for_Second_Language_Learning_Purposes`.

Sarah E Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 523–530. Association for Computational Linguistics.

Thomas M Segler. 2007. *Investigating the selection of example sentences for unknown target words in ICALL reading texts for L2 German*. PhD Thesis. University of Edinburgh.

Wade Shen, Jennifer Williams, Tamas Marius, and Elizabeth Salesky. 2013. A language-independent approach to automatic text difficulty assessment for second-language learners. In *Proceedings of the 2nd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 30–38. Association for Computational Linguistics.

Ajay Kumar Tanwani, Jamal Afridi, M Zubair Shafiq, and Muddassar Farooq. 2009. Guidelines to select machine learning scheme for classification of biomedical datasets. In *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, pages 128–139. Springer.

Sowmya Vajjala and Detmar Meurers. 2012. 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*, pages 163–173. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2014. 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.

Elena Volodina and Sofie Johansson Kokkinakis. 2013. Compiling a corpus of CEFR-related texts. In *Proceedings of the Language Testing and CEFR conference, Antwerpen, Belgium, May 27-29, 2013*.

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, Richard Johansson, and Sofie Johansson Kokkinakis. 2012b. Semi-automatic selection of best corpus examples for Swedish: Initial algorithm evaluation. In *Workshop on NLP in Computer-Assisted Language Learning. Proceedings of the SLTC 2012 workshop on NLP for CALL. Linköping Electronic Conference Proceedings*, volume 80, pages 59–70.

Elena Volodina, Dijana Pijetlovic, Ildikó Pilán, and Sofie Johansson Kokkinakis. 2013. Towards a gold standard for Swedish CEFR-based ICALL. In *Proceedings of the Second Workshop on NLP for Computer-Assisted Language Learning. NEALT Proceedings Series 17. Nodalida 2013, Oslo, Norway.*