Accuracy scores using other learning algorithms were significantly lower (see Table 4), therefore, we report only the results of the logistic regression classifier in the subsequent sections.

Table 4. Accuracy scores (in %) for other learning algorithms.

Type	Nr	Perceptron	SMO	J48
LEX	11	**77.4**	42.1	55
ALL	61	62.2	52.7	50.5

Instead of classification, some readability studies (e.g. [11, 15]) employed linear regression for this task. For a better comparability, we applied also a linear regression model to our data which yielded a correlation of 0.8 and an RMSE of 0.65.

To make sure that our system was not biased towards the majority classes B1 and B2, we inspected the confusion matrix (Table 5) after classification using ALL. We can observe from Table 5 that the system performs better at A1 and C1 levels, where confusion occurred only with adjacent classes. Similar to the findings in [14] for French, classes in the middle of the scale were harder to distinguish. Most misclassifications in our material occurred at A2 level (23%) followed by B1 and B2 level, (20% and 17% respectively).

Table 5. Confusion matrix for feature set ALL at document level.

Predictions						
A1	A2	B1	B2	C1		
37	12	0	0	0	A1	L
12	121	18	5	1	A2	a
4	11	206	24	13	B1	b
0	5	21	238	24	B2	e
0	0	0	12	103	C1	l

To establish the external validity of our approach, we tested it on a subset of LÄSBART [5], a corpus of Swedish easy-to-read (ETR) texts previously employed for Swedish L1 readability studies [5, 18]. We used 18 fiction texts written for children between ages nine to twelve, half of which belonged to the ETR category and the rest were unsimplified. Our model generalized well to unseen data, it classified all ETR texts as B1 and all ordinary texts as C1 level, thus correctly identifying in all cases the relative difference in complexity between the documents of the two categories.

Although a direct comparison with other studies is difficult because of the target language, the nature of the datasets and the number of classes used, in

terms of absolute numbers, our model achieves comparable performance with the state-of-the-art systems for English[10, 13]. Other studies for non-English languages using CEFR levels include: [14], reporting 49.1% accuracy for a French system distinguishing six classes; and [15] achieving 29.7% accuracy on a smaller Portuguese dataset with five levels.

4.3 Sentence-Level Experiments

After building good classification models at document level, we explored the usability of our approach at the sentence level. Sentences are particularly useful in Computer-Assisted Language Learning (CALL) applications, among others, for generating sentence-based multiple choice exercises, e.g. [26], or vocabulary examples [27]. Furthermore, multi-class readability classification of sentence-level material intended for second language learners has not been previously investigated in the literature.

With the same methodology (section 4.1) and feature set (section 3) used at the document level, we trained and tested classification models based on the sentence-level data (see section 2). The results are shown in Table 6.

Table 6. Sentence-level classification results.

Type	Nr	Acc (%)	F	RMSE
Majority	-	40.2	0.23	0.49
LIX	1	41.4	0.3	0.38
Lex	11	56.8	0.53	0.33
All	61	**63.4**	**0.63**	0.31

Although the majority baseline in the case of sentences was 7% higher than the one for texts (Table 3), the classification accuracy for sentences using all features was only 63.4%. This is a considerable drop (-18%) in performance compared to the document level (81.3% accuracy). It is possible that the features did not capture differences between the sentences because the amount of context is more limited on the fine-grained level. It is interesting to note that, although there was no substantial performance difference between Lex and All at a document level, the model with all the features performed 7% better at sentence level.

Most misclassifications occurred, however, within a distance of one class only, thus the adjacent accuracy of the sentence-level model was still high, 92% (see Table 7). Predictions were noticeably more accurate for classes A1, A2 and B1 which had a larger number of instances.

In the next step, we applied the sentence-level model on the document-level data to explore how homogeneous texts were in terms of the CEFR level of the sentences they contained. Figure 1 shows that texts at each CEFR level contain

Table 7. Confusion matrix for feature set ALL at sentence level.

Predictions						
A1	A2	B1	B2	C1		
371	123	9	2	0	A1	L
120	541	78	11	4	A2	a
27	136	212	23	10	B1	b
8	34	39	30	13	B2	e
0	18	21	9	35	C1	l

a substantial amount of sentences of the same level of the whole text, but they also include sentences classified as belonging to other CEFR levels.

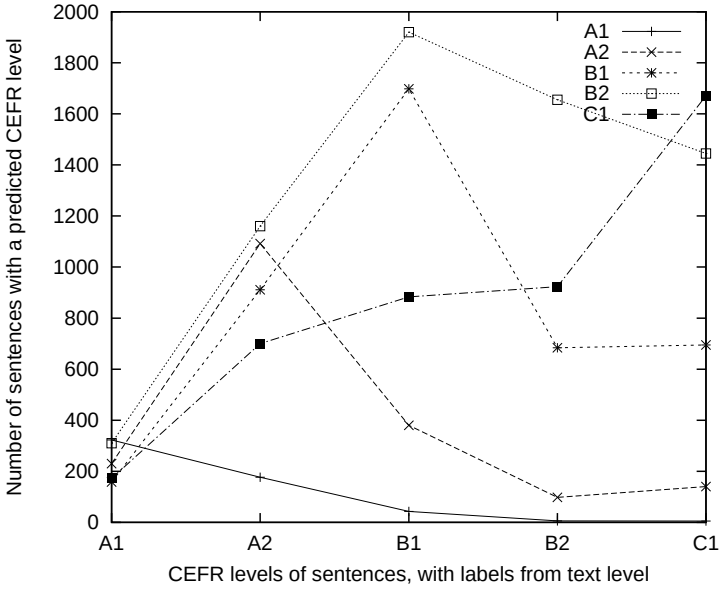

Fig. 1. Distribution of sentences per CEFR level in the document-level data.

Finally, as in the case of the document-level analysis, we tested our sentence-level model also on an independent dataset (SENREAD), a small corpus of sentences with gold-standard CEFR annotation. This data was created during a user-based evaluation study [28] and it consists of 196 sentences from generic corpora, i.e. originally not L2 learner-focused corpora, rated as being suitable at B1 or being at a level higher than B1. We used this corpus along with the judgments of the three participating teachers. Since SENREAD had only two