Table 1. The distribution of items per CEFR level in the datasets.

	Document level				Sentence level	
CEFR	Books	Publ.	Texts	Mean nr. sent	Books	Sentences
A1	4	3	49	14.0	4	505
A2	4	3	157	13.8	4	754
B1	5	3	258	17.9	4	408
B2	4	3	288	26.6	3	124
C1	2	2	115	42.1	1	83
Total	12	4	867	-	4	1874

sentence-level analysis, instead of the ratio of number of tokens to the number of sentences in the text, we considered the number of tokens in one sentence.

Lexical (Lex): Similar to [9], we used information from the Kelly list [22], a lexical resource providing a CEFR level and frequencies per lemma based on a corpus of web texts. Thus, this word list is entirely independent from our dataset. Instead of percentages, we used *incidence scores* (IncSc) per 1000 words to reduce the influence of sentence length on feature values. The IncSc of a category was computed as 1000 divided by the number of tokens in the text or sentence multiplied by the count of the category in the sentence. We calculated the IncSc of words belonging to each CEFR level (#6 - #11). In features #12 and #13 we considered *difficult* all tokens whose level was above the CEFR level of the text or sentence. We computed also the IncSc of tokens not present in the Kelly list (#14), tokens for which the lemmatizer did not find a corresponding lemma form (# 15), as well as average log frequencies (#16).

Morphological (Morph): We included the variation (the ratio of a category to the ratio of lexical tokens - i.e. nouns, verbs, adjectives and adverbs) and the IncSc of all lexical categories together with the IncSc of punctuations, particles, sub- and conjunctions (#34, #51). Some additional features, using insights from L2 teaching material [21], captured fine-grained inflectional information such as the IncSc of neuter gender nouns and the ratio of different verb forms to all verbs (#52 - #56). Instead of simple type-token ratio (TTR) we used a bilogarithmic and a square root TTR as in [4]. Moreover, nominal ratio [5], the ratio of pronouns to prepositions [14], and two *lexical density* features were also included: the ratio of lexical words to all non-lexical categories (#48) and to all tokens (#49). Relative structures (#57) consisted of relative adverbs, determiners, pronouns and possessives.

Syntactic (Synt): Some of these features were based on the length (depth) and the direction of dependency arcs[5] (#17 - #21). We complemented this, among others, with the IncSc of relative clauses in clefts[6] (#26), and the IncSc of pre-and postmodifiers (e.g. adjectives and prepositional phrases) [5].

[5] The tags were obtained with the MaltParser [23].

[6] Sentences that begin with a constituent receiving particular focus, followed by a relative clause. E.g.: It is John (whom) Jack is waiting for.

Semantic (SEM): Features based on information from SALDO [24], a Swedish lexical-semantic resource. We used the average number of senses per token as in [9] and included also the average number of noun senses per nouns. Once reliable word-sense disambiguation methods become available for Swedish, additional features based on word senses could be taken into consideration here.

The complete set of 61 features is presented in Table 2. Throughout this paper we will refer to the machine learning models using this set of features, unless otherwise specified. Features for both document- and sentence-level analyses were extracted for each sentence, the values being averaged over all sentences in the text in the document-level experiments to ensure comparability.

Table 2. The complete feature set.

Nr.	Feature Name	Nr.	Feature Name
	Length-based		*Morphological*
1	Sentence length	30	Modal verbs to verbs
2	Average token length	31	Particle INCSC
3	Extra-long words	32	3SG pronoun INCSC
4	Number of characters	33	Punctuation INCSC
5	LIX	34	Subjunction INCSC
	Lexical	35	S-verb INCSC
6	A1 lemma INCSC	36	S-verbs to verbs
7	A2 lemma INCSC	37	Adjective INCSC
8	B1 lemma INCSC	38	Adjective variation
9	B2 lemma INCSC	39	Adverb INCSC
10	C1 lemma INCSC	40	Adverb variation
11	C2 lemma INCSC	41	Noun INCSC
12	Difficult word INCSC	42	Noun variation
13	Difficult noun and verb INCSC	43	Verb INCSC
14	Out-of-Kelly INCSC	44	Verb variation
15	Missing lemma form INCSC	45	Nominal ratio
16	Avg. Kelly log frequency	46	Nouns to verbs
	Syntactic	47	Function word INCSC
17	Average dependency length	48	Lexical words to non-lexical words
18	Dependency arcs longer than 5	49	Lexical words to all tokens
19	Longest dependency from root node	50	Neuter gender noun INCSC
20	Ratio of right dependency arcs	51	Con- and subjunction INCSC
21	Ratio of left dependency arcs	52	Past participles to verbs
22	Modifier variation	53	Present participles to verbs
23	Pre-modifier INCSC	54	Past verbs to verbs
24	Post-modifier INCSC	55	Present verbs to verbs
25	Subordinate INCSC	56	Supine verbs to verbs
26	Relative clause INCSC	57	Relative structure INCSC
27	Prepositional complement INCSC	58	Bilog type-token ratio
	Semantic	59	Square root type-token ratio
28	Avg. nr. of senses per token	60	Pronouns to nouns
29	Noun senses per noun	61	Pronouns to prepositions

4 Experiments and Results

4.1 Experimental Setup

We explored different classification algorithms for this task using the machine learning toolkit WEKA [25]. These included: (1) a multinomial logistic regression model with ridge estimator, (2) a multilayer perceptron, (3) a support vector machine learner, Sequential Minimal Optimization (SMO), and (4) a decision tree (J48). For each of these, the default parameter settings have been used as implemented in WEKA.

We considered classification accuracy, F-score and Root Mean Squared Error (RMSE) as evaluation measures for our approach. We also included a confusion matrix, as we deal with a dataset that is unbalanced across CEFR levels. The scores were obtained by performing a ten-fold Cross-Validation (CV).

4.2 Document-Level Experiments

We trained document-level classification models, comparing the performance between different subgroups of features. We had two baselines: a majority classifier (Majority), with B2 as majority class, and the LIX readability score. Table 3 shows the type of subgroup (*Type*), the number of features (*Nr*) and three evaluation metrics using logistic regression.

Table 3. Document-level classification results.

Type	Nr	Acc (%)	F	RMSE
Majority	-	33.2	0.17	0.52
LIX	1	34.9	0.22	0.38
Lex	11	**80.3**	**0.80**	0.24
All	61	**81.3**	**0.81**	0.27

Not only was accuracy very low with LIX, but this measure also classified 91.6% of the instances as B2 level. Length-based, semantic and syntactic features in isolation showed similar or only slightly better performance than the baselines, therefore we excluded them from Table 3. Lexical features, however, had a strong discriminatory power without an increase in bias towards the majority classes. Using this subset of features only, we achieved approximately the same performance (0.8 F) as with the complete set of features, All (0.81 F). This suggests that lexical information alone can successfully distinguish the CEFR level of course book texts at the document level. Using the complete feature set we obtained 81% accuracy and 97% *adjacent accuracy* (when misclassifications to adjacent classes are considered correct). The same scores with lexical features (Lex) only were 80.3% (accuracy) and 98% (adjacent accuracy).