# Investigating the importance of linguistic complexity features across different datasets related to language learning

**Ildikó Pilán**
Språkbanken, University of Gothenburg
Sweden
ildiko.pilan@gu.se

**Elena Volodina**
Språkbanken, University of Gothenburg
Sweden
elena.volodina@gu.se

## Abstract

We present the results of our investigations aiming at identifying the most informative linguistic complexity features for classifying language learning levels in three different datasets. The datasets vary across two dimensions: the size of the instances (texts vs. sentences) and the language learning skill they involve (reading comprehension texts vs. texts written by learners themselves). We present a subset of the most predictive features for each dataset, taking into consideration significant differences in their per-class mean values and show that these subsets lead not only to simpler models, but also to an improved classification performance. Furthermore, we pinpoint fourteen central features that are good predictors regardless of the size of the linguistic unit analyzed or the skills involved, which include both morpho-syntactic and lexical dimensions.

## 1 Introduction

Linguistic complexity, especially in cross-linguistic studies, is often approached in absolute terms, describing complexity as a property of a linguistic system in terms of e.g. number of contrastive sounds. In this paper, however, we investigate a *relative* type of linguistic complexity from a cognitive perspective, our focus being the ability of L2 learners to process or produce certain linguistic elements in writing at different stages of proficiency. We operationalize the term *linguistic complexity* as the set of lexico-semantic, morphological and syntactic characteristics reflected in texts (or sentences) that determine the magnitude of the language skills and competences required to process or produce them. In this work, we use linguistic complexity analysis as a means to predict second language learning (L2) levels. The scale of learning (*proficiency*) levels adopted here is the CEFR, the Common European Framework of Reference for Languages (Council of Europe, 2001) which proposes a six-point scale of proficiency levels: from A1 (beginner) to C2 (advanced) level.

Large corpora in the language learning domain are rather scarce due to either copy-right issues, privacy reasons or the need for digitizing them. For the Swedish language, a number of resources have become available recently (Volodina et al., 2014; Volodina et al., 2016b), which, although somewhat small in size, encompass texts involving different skills and CEFR levels. This allows for investigations about the similarities and differences between linguistic complexity observable at different proficiency levels for different skill types, namely *receptive* skills, required when learners process passages produced by others and *productive* skills, when learners produce the texts themselves. We perform linguistic complexity analyses across two different dimensions: the type of learner skills involved when dealing with the texts and the size of the linguistic context investigated. In the latter case, we carry out experiments both at the text and at the sentence level.

Throughout the years, a large number of linguistic features related to complexity has been proposed. Typically, out of the features suggested for a specific task some are more useful than others. Eliminating redundant features can result in simpler and improved models that are not only faster, but might also generalize better on unseen data (Witten et al., 2011, 308). Such selection can also contribute to understand further the main factors playing role in linguistic complexity, which can be a useful means for

*Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 49–58
Santa Fe, New Mexico, USA, August 25, 2018.

determining whether non-native speakers can understand or produce certain linguistic input at different learning levels. In this paper, we investigate therefore the importance of individual linguistic complexity features for predicting proficiency levels across different L2 datasets. The two main research questions we investigate are: (i) Which linguistic complexity features are most useful for determining proficiency levels for each L2 dataset? (ii) Are there features that are relevant regardless of the context size and the type of skill considered? Our contributions include, on the one hand, a subset of the most informative features for each dataset whose use leads to improved classification results. On the other hand, we identify some lexical, morphological and syntactic features that are good indicators of complexity across all three datasets, namely, reading comprehension texts, essays and sentences.

In Section 2, we provide an overview of previous work related to linguistic complexity analysis, followed by the description of our datasets in Section 3. In Section 4, we present the set of features used and highlight their relevance for modeling linguistic complexity in the L2 context. We then describe our experiments and their results in Section 5, presenting the most informative features and their effect on classification performance. Finally, we conclude our results and outline future work in Section 6.

## 2 Previous literature on linguistic complexity for predicting L2 levels

**Expert-written (receptive) texts** In the L2 context, specific scales reflecting progress in language proficiency have been proposed. One such scale is the CEFR, introduced in section 1. An alternative to the CEFR is the 7-point scale of the Interagency Language Roundtable (ILR), common in the United States. In Table 1, we provide an overview of studies targeting L2 receptive complexity and compare the target language, the type and amount of training data and the methods used. The studies are ordered alphabetically based on the target language of the linguistic complexity analysis. We only include previous work here that shares the following characteristics: (i) texts rather than single sentences are the unit of analysis; (ii) receptive linguistic complexity is measured; and (iii) NLP tools are combined with machine learning algorithms. Under dataset size, we report the number of texts used (except for Heilman et al. (2007)), where whole books were employed), followed by the number of tokens in parenthesis when available.

| Study | Target language | CEFR | Dataset size in # texts | Text type | # levels | Method |
|---|---|---|---|---|---|---|
| Salesky and Shen (2014) | Arabic, Dari English, Pashto | No | 4 × 1400 | Non-L2 | 7 | Regr. |
| Sung et al. (2015) | Chinese | Yes | 1578 | L2 | 6 | Classif. |
| Heilman et al. (2007) | English | No | 4 books (200,000) | L2 | 4 | Regr. |
| Huang et al. (2011) | English | No | 187 | Both | 6 | Regr. |
| Xia et al. (2016) | English | Yes | 331 | L2 | 5 (A2-C2) | Both |
| Zhang et al. (2013) | English | No | 15 | Non-L2 | 1-10 | Regr. |
| François and Fairon (2012) | French | Yes | 1852 (510,543) | L2 | 6 | Classif. |
| Branco et al. (2014) | Portuguese | Yes | 110 (12,673) | L2 | 5 (A1-C1) | Regr. |
| Curto et al. (2015) | Portuguese | Yes | 237 (25,888) | L2 | 5 (A1-C1) | Classif. |
| Karpov et al. (2014) | Russian | Yes | 219 | Both | 4 (A1-B1, C2) | Classif. |
| Reynolds (2016) | Russian | Yes | 4689 | Both | 6 | Classif. |
| Pilán et al. (2016) | Swedish | Yes | 867 | L2 | 5 (A1-C1) | Both |

Table 1: An overview of studies on L2 receptive complexity.

CEFR-based studies have been more commonly treated as a classification problem, a popular choice of classifier being support vector machines (SVM). A particular aspect distinguishing Xia et al. (2016) from the rest of the studies mentioned in Table 1 is the idea of using L1 data to improve the classification of L2 texts. For the sake of comparability, the information in Table 1 describes only the experiments using the L2 data reported in this study. The state-of-the-art performance reported for the CEFR-based classification described in the studies included in Table 1 ranges between 75% and 80% accuracy (Curto et al., 2015; Sung et al., 2015; Xia et al., 2016; Pilán et al., 2016a).

A large number of features have been proposed and tested in this context. Count-based measures (e.g. sentence and token length, type-token ratio) and syntactic features (e.g. dependency length) have been confirmed to be influencing factors in L2 complexity (Curto et al., 2015; Reynolds, 2016). Lexical information based on either n-gram models (Heilman et al., 2007) or frequency information from word lists (François and Fairon, 2012; Reynolds, 2016; Salesky and Shen, 2014) and Google search results (Huang et al., 2011) has proven to be, however, one of the most predictive dimensions. Heilman et al. (2007) found that lexical features outperform grammatical ones, which, although more important for L2 than L1 complexity, still remain less predictive for L2 English complexity. Nevertheless, the authors mention that this may depend on the morphological richness of a language. Reynolds (2016), in fact, finds that morphological features are among the most influential ones for L2 Russian texts.

**Learner-written (productive) texts** Similarly to L2 texts targeting reading skills, also texts produced by L2 learners manifest varying degrees of complexity at different stages of proficiency. Typically however, receptive linguistic complexity is somewhat higher than its productive counterpart for a learner at a given CEFR level (Barrot, 2015). Previous studies aiming at classifying CEFR levels in learner-written texts include Hancke and Meurers (2013) for L2 German and Vajjala and Lõo (2014) for L2 Estonian. The most predictive features for L2 German include lexical and morphological features. Morphological features (e.g. amount of distinct cases used) are also among the most informative ones for L2 Estonian at all L2 development stages. A fundamental difference between assessing receptive and productive texts is that, while receptive texts are expected to be relatively error free, the latter ones typically contain a varying amount of L2 errors, which have also been used to inform features. Errors are usually counted based on the output of a spell checker (Hancke and Meurers, 2013; Tack et al., 2017) or by using hand-crafted rules (Tack et al., 2017).

**Smaller linguistic units** Besides the text-level analyses in Table 1, studies targeting smaller units also appear in the literature. Linguistic complexity in single sentences from an L2 perspective has been explored in Karpov et al. (2014) and in Pilán et al. (2016a). Both studies are CEFR-related, but rather than classifying sentences into individual CEFR levels, a binary distinction is made (below or at B1 level vs. above B1). In Pilán et al. (2016a), we report 63% accuracy for a 5-way CEFR level classification of Swedish coursebook sentences. As for productive complexity, research on the automatic assessment of short answers to open-ended questions in terms of using CEFR has been investigated in Tack et al. (2017) for L2 English. The authors proposed an ensemble method consisting of integrating the votes of a number of traditional classification methods into a single prediction. Sentence and word length, lexical features and information about the age of acquisition of words were found especially predictive.

## 3 Datasets

### 3.1 Text-level datasets

We used two L2 Swedish corpora consisting of texts in our experiments: SweLL (Volodina et al., 2016b) comprised of essays written by L2 learners and COCTAILL (Volodina et al., 2014) containing L2 coursebooks authored or adapted by experts for L2 learners. The SweLL corpus consists of essays produced by adult learners of L2 Swedish on a variety of topics (TEXT-E). From the coursebook corpus, we only include whole texts meant for reading comprehension practice (TEXT-R) since the linguistic annotation of other coursebook elements (e.g. gap-filling exercises) may be prone to automatic linguistic annotation errors. These two corpora cover five CEFR levels (A1 to C1). Each SweLL essay has been assigned a CEFR level by teachers. For reading texts, CEFR levels were derived from the level of the lesson (chapter) they occur in. It is worth mentioning that these two corpora are *independent* from each other, i.e. the essays written by the learners are not based on, or inspired by, the reading passages. The distribution of texts per type and CEFR level in the datasets is shown in Table 2. The total number of tokens in the coursebook-based dataset was 289,312, while in the learner essay data it was 43,033.

### 3.2 A teacher-evaluated dataset of sentences

At the sentence level, we use a small dataset[1] (SENT) based on the user evaluation of a corpus example selection system, HitEx, which we described in detail in Pilán et al. (2016b). HitEx aims at identifying

[1] The dataset is available at `https://github.com/IldikoPilan/sent_cefr`.

sentences from corpora suitable as exercise items. The sentences in this dataset have been automatically assessed for their CEFR level and have been filtered for their well-formedness, independence from the rest of their textual context and some additional lexical and structural criteria (e.g. abbreviations, interrogative form) using HitEx. Out of the original 330 sentences from the evaluation material, we only included in this dataset the subset of sentences: (i) that were found overall suitable (with an evaluation score $>=$ 2.5 out of 4); and (ii) where a majority of teachers agreed with the CEFR level assigned automatically by HitEx. This subset was complemented with 90 sentences for the otherwise insufficiently represented A1 level from the COCTAILL corpus. Only individually occurring sentences in lists and non-gapped exercises were considered, thus these are not a subset of the text-level dataset described above. The distribution of sentences per CEFR level in the dataset is presented in Table 2. The total number of tokens in the dataset is 4,060.

| **Writer** | **Unit** | **A1** | **A2** | **B1** | **B2** | **C1** | **Total** |
|---|---|---|---|---|---|---|---|
| **Learner** | **Texts** | 16 | 83 | 75 | 74 | 88 | **336** |
| **Expert** | **Texts** | 49 | 157 | 258 | 288 | 115 | **867** |
| **Expert** | **Sentences** | 98 | 82 | 58 | 92 | 45 | **375** |

Table 2: CEFR-level annotated Swedish datasets.

All three corpora are equipped also with automatic linguistic annotation which includes lemmatization, part-of-speech (POS) tagging and dependency parsing based on the *Sparv*[2] pipeline.

## 4 A flexible feature set for linguistic complexity analysis

In this section, we provide a detailed description of the set of features used and relate them to cognitive aspects of linguistic complexity. The feature set is “flexible” in the sense that it can be applied to different types of L2 data and units of analysis (e.g. texts or sentences) since it does not incorporate text-level features (e.g. discourse-related aspects) or learner language specific ones (e.g. L2 error features). The feature set is comprised of 61 features in total, which we have previously used for CEFR classification experiments also in Pilán et al. (2016c). Table 3 shows the complete feature set divided into five subcategories based on the type of NLP tools and resources used: *count-based*, *lexical*, *morphological*, *syntactic* and *semantic*.

### 4.1 Count-based features

The feature set includes seven indicators that are based on simple counts or traditional readability measures. One such measure for Swedish is *LIX* (*Läsbarhetsindex* ‘Readability index’) proposed in Björnsson (1968). LIX combines the sum of the average number of words per sentence in the text and the percentage of tokens longer than six characters. Sentence length is measured both as the number of tokens and that of characters. Sentence length can indicate syntactic difficulty and it can be a sign of e.g. multiple clauses or larger noun phrases. Average token ($T$) length is computed based on the number of characters. Extra-long words, i.e. tokens longer than 13 characters, are also counted since compounding, frequent in Swedish, can result in particularly long words (Heimann Mühlenbock, 2013). Type-token ratio (TTR), the ratio of unique tokens to all tokens, is an indicator of lexical richness (Graesser et al., 2004). A bi-logarithmic and a square root TTR are used which decrease the effect of text and sentence length (Vajjala and Meurers, 2012).

### 4.2 Word-list based lexical features

Besides richness, the frequency of words also influences lexical complexity as repeated exposure facilitates their processing (Graesser et al., 2004). Frequency information is collected from the KELLY list (Volodina and Kokkinakis, 2012), based on web texts.

[2] `https://spraakbanken.gu.se/sparv/`

| COUNT | SYNTACTIC | MORPHOLOGICAL |
|---|---|---|
| Sentence length | Avg. DepArc length | Function W IncSc |
| Avg token length | DepArc Len > 5 | Particle IncSc |
| Extra-long token | Max length DepArc | 3SG pronoun IncSc |
| Nr characters | Right DepArc Ratio | Punctuation IncSc |
| LIX | Left DepArc Ratio | Subjunction IncSc |
| Bilog TTR | Modifier variation | PR to N |
| Square root TTR | Pre-modifier IncSc | PR to PP |
| LEXICAL | Post-modifier IncSc | Relative structure IncSc |
| Avg KELLY log freq | Subordinate IncSc | S-V IncSc |
| A1 lemma IncSc | Relative clause IncSc | S-V to V |
| A2 lemma IncSc | PP complement IncSc | ADJ IncSc |
| B1 lemma IncSc | MORPHOLOGICAL | ADJ variation |
| B2 lemma IncSc | Neuter N IncSc | ADV IncSc |
| C1 lemma IncSc | CJ + SJ IncSc | ADV variation |
| C2 lemma IncSc | Past PC to V | N IncSc |
| Difficult W IncSc | Present PC to V | N variation |
| Difficult N&V IncSc | Past V to V | V IncSc |
| OOV IncSc | Supine V to V | V variation |
| No lemma IncSc | Present V to V | Lex T to Nr T |
| SEMANTIC | Nominal ratio | Lex T to non-lex T |
| Avg senses per token | N to V | |
| N senses per N | Modal V to V | |

Table 3: Feature set for linguistic complexity assessment in L2 data.

Instead of n-grams, weakly lexicalized features are employed to increase the generalizability of the models on unseen data. Each token is represented by its corresponding CEFR level. Unlike in Pilán et al. (2016c), where we employed KELLY, the per-token CEFR level information is retrieved here from two word lists compiled based on the L2 corpora described in Section 3. To guarantee the independence of the word lists from the datasets, we use SweLLex (Volodina et al., 2016a), a frequency list based on the learner essays when classifying CEFR levels in coursebook texts and SVALex (François et al., 2016), containing frequencies from coursebooks for making predictions on the essays. For sentences, SVALex has been used since it is independent from the dataset, but both reflect receptive linguistic complexity. Frequency distributions in these lists have been mapped to single CEFR levels based on the difference in per-level normalized frequency between adjacent levels as described in Alfter et al. (2016).

Instead of absolute counts, a normalized value, an *incidence score* (IncSc = $\frac{1000}{N_t} \times N_c$) is used to reduce the influence of sentence length, where $N_t$ is the total number of tokens and $N_c$ is the count of a certain category of tokens in the text or sentence (Graesser et al., 2004). The IncSc of *difficult* tokens is also computed, that is, tokens above a certain reference CEFR level, which can be the level of an L2 learner writing a text or whom the text would be presented to as reading material. This value is also computed separately for nouns and verbs, since these are crucial for conveying meaning. Moreover, the IncSc of tokens not present in the L2 word lists, i.e. out-of-vocabulary words (*OOV* IncSc) is also considered as well as the IncSc of non-lemmatized tokens (*No lemma* IncSc).

## 4.3 Morphological features

*Morphological features* include not only IncSc of different morpho-syntactic categories, but also variational scores, i.e. the ratio of a category to the ratio of *lexical* tokens: nouns (N), verbs (V), adjectives (ADJ) and adverbs (ADV). Some specific features for L2 Swedish are the ratio of different verb forms to verbs which are typically introduced at varying stages of L2 learning. *S-verbs* (*S-VB*) are a group of Swedish verbs ending in *-s* that are peculiar in terms of morphology and semantics. They indicate either reciprocity, a passive construction or are *deponent* verbs, i.e. verbs active in meaning, but passive in form. Neuter gender nouns are also considered since they can indicate the abstractness of a concept (Graesser et al., 2004). Among relative structures relative adverbs, determiners, pronouns and possessives are counted. *Nominal ratio* (Hultman and Westman, 1977) corresponds to the ratio of nominal categories, i.e. nouns, prepositions (PP) and participles to the ratio of verbal categories, namely pronouns (PR), adverbs, and verbs. Its simplified version is the ratio of nouns to verbs, and it is meant to measure the information load of a text or reveal its genre (e.g. spoken vs. news text). A higher value corresponds to higher degrees of complexity and a more elaborate genre.

INCSC for punctuation marks as well as sub- and conjunctions (SJ, CJ) are also computed since their presence in larger quantities can indicate a more complex syntactic structure. Particles can change the meaning of verbs considerably, similarly to English phrasal verbs (Heimann Mühlenbock, 2013). The INCSC of the third person singular (3SG) pronoun inspired by Zhang et al. (2013) is also included since this is often used referentially, which can further increase the difficulty of processing.

### 4.4 Syntactic and semantic features

Syntactic aspects are related to readers' working memory load when processing sentences which can be increased by ambiguity or embedded constituents (Graesser et al., 2004). Here, the average length (depth) of dependency arcs (*DepArc*) and their direction is considered. Relative clauses, pre- and post-modifiers (e.g. adjectives and prepositional phrases), prepositional complements as well as subordinates, commonly used in previous research on linguistic complexity (Heimann Mühlenbock, 2013; Schwarm and Ostendorf, 2005), are also counted.

The two semantic features included quantify available word senses per lemma based on the SALDO lexicon (Borin et al., 2013). Both the average number of senses per token and the average number of noun senses per noun are considered. Polysemous words can be demanding for readers as they need to be disambiguated for a full understanding of the sentence (Graesser et al., 2004).

## 5 Cross-dataset feature selection experiments

In this section, we describe the results of our feature selection experiments on the three datasets presented in Section 3. These experiments differ from the ones we described previously in Pilán et al. (2016a) and Pilán et al. (2016c) in a number of respects. In this work, the worth of individual features is evaluated rather than that of the complete set of features or groups of features. Moreover, as mentioned in section 4, most lexical features are based on L2 word lists rather than KELLY.

### 5.1 Experimental setup

We use 85% of each dataset for identifying the most informative features (DEV). The reported classification results using this part of the data are based on a stratified 5-fold cross-validation setup, that is, the original distribution of instances per CEFR level in the dataset has been preserved in all folds. We evaluated the generalizability of the selected subset of features on the remaining 15% of the data (TEST). As learning algorithm for these models, we used *LinearSVC* as implemented in scikit-learn (Pedregosa et al., 2011), which has been successfully applied in recent years in a number of NLP areas.

### 5.2 Feature selection method

As a pre-processing step before training our classifiers, we used a *univariate feature selection* method, also available in scikit-learn, to identify the most informative features scored with *analysis of variance* (ANOVA). This feature selection method is suitable for multi-class problems, it is independent of the learning method used and it has been previously adopted for NLP tasks, e.g. by Carbon et al. (2014). ANOVA is a statistical test that can be used to measure how strong the relationship between each feature and the output class is (CEFR levels in our case). It relies on *F-tests*, which can be employed to score features based on significant differences in their per-class mean values. To detect these differences indicating dependencies, first, the *variance*, i.e. the dispersion of the data in terms of its distance from the mean, is measured both *within* and *between* classes for each feature. Then, the F-statistic can be computed as the ratio of the variance between class means and the variance within a class.

### 5.3 Results

The results of the models with and without feature selection in terms of accuracy and $F_1$ are presented in Table 4.

| **Data** | **Features** | **SENT** | | **TEXT-R** | | **TEXT-E** | |
|---|---|---|---|---|---|---|---|
| | | Acc | $F_1$ | Acc | $F_1$ | Acc | $F_1$ |
| DEV | ALL | 0.62 | 0.61 | 0.68 | 0.68 | 0.73 | 0.71 |
| DEV | K-BEST | 0.73 | 0.71 | 0.70 | 0.70 | 0.81 | 0.81 |
| TEST | K-BEST | 0.81 | 0.79 | 0.73 | 0.73 | 0.84 | 0.82 |
| Number of K-BEST | | 21 | | 54 | | 24 | |

Table 4: Accuracy with feature selection across datasets.

Reducing the complete set of features to the subset of the most informative ones improved the classification results for all datasets. The most substantial boost (+0.11 accuracy) was obtained for sentences. The models with selected features generalized well also on the held-out test sets. Moreover, while for SENT and TEXT-E only about one third of the features have been selected, almost all features were included in the $k$ number of best ones for TEXT-R. The selected features ranked based on ANOVA are presented in Table 5. For TEXT-R, features with low importance are not listed separately. These are only indicated when they overlap with a feature selected by the other models (with a rank $> 24$).

Fourteen features were among the most informative ones across all three datasets, which are highlighted in bold in Table 5. One such feature was the count-based measure of square root TTR, thus it seems that a varied way of expression, through e.g. the use of synonyms, is a good indicator of linguistic complexity in the L2 context. Among the word-list based lexical features, besides the proportion of difficult lexica, the amount of tokens at the extremes of the CEFR scale, namely the lowest, A1 level and the advanced, C1 level (the highest available in our L2 lists) were also useful predictors. Interestingly, two out of the three strong indicators of L2 English essay quality identified in Crossley and McNamara (2011) were lexical diversity, closely related to our Square root TTR feature, and lexical frequency, based on the same type of information as our word-list features. Lexical variation in terms of TTR as well as verb variation were also found highly predictive for L2 Estonian learner texts (Vajjala and Lõo, 2014). These findings indicate the predictive strength of these features across languages. Furthermore, syntactic features relative to the length of dependency arcs and verb-related morphological features (e.g. INCSC of participles and s-verbs) were among the $k$-best for all datasets. Such verb forms are, in fact, typically introduced explicitly to L2 learners at higher CEFR levels (Fasth and Kannermark, 1997). The amount of punctuation and particles was also indicative of complexity. The former can, for example, indicate clause boundaries and hence more complex sentences. Particles, on the other hand, can be challenging for language learners, since they alter the meaning of verbs.

For the two datasets related to receptive skills, SENT and TEXT-R, a number of count features were strongly predictive. Unlike for TEXT-E, sentence length in terms of both the number of tokens and the number of characters were highly informative for determining receptive complexity. Although the proportion of lexical tokens to all tokens was not informative at the sentence level, it proved to be a good indicator of linguistic complexity at the text level. The traditional readability measure, LIX was informative only for TEXT-R, which could be explained by the fact that this dataset was the most similar to the intended use of LIX, namely determining readability at the text level. On the other hand, the other traditional formula, nominal ratio, was more useful across datasets, especially in its simplified version (*N to V*). It would be useful to investigate further whether this also depends on a difference in text genre.

A limitation of our study is the relatively small size of our datasets, which is especially true in the case of the A1 level learner essays. Considering the difficulties in having access to similar types of L2 data, and the extension of our experiments to cross-dataset observations, the results could still provide valuable insights for teaching experts and members of the NLP community targeting similar tasks.

| Feature name | Rank | | |
|---|---|---|---|
| | Sent | Text-R | Text-E |
| Nr characters | 1 | 4 | - |
| **Square root TTR** | 2 | 7 | 9 |
| **A1 lemma IncSc** | 3 | 3 | 2 |
| **Punctuation IncSc** | 4 | 11 | 12 |
| Sentence length | 5 | 5 | - |
| **Relative clause** | 6 | > 24 | 8 |
| **Difficult N&V IncSc** | 7 | 1 | 1 |
| **Avg. DepArc length** | 8 | 10 | 14 |
| **Max length DepArc** | 9 | 6 | 13 |
| Bilog TTR | 10 | 24 | - |
| DepArc Len > 5 | 11 | 8 | - |
| S-V IncSc | 12 | > 24 | - |
| **Present PC to V** | 13 | 18 | 17 |
| **Past PC to V** | 14 | > 24 | 18 |
| **Particle IncSc** | 15 | > 24 | 16 |
| **V variation** | 16 | 15 | 10 |
| **Difficult W IncSc** | 17 | 2 | 4 |
| V IncSc | 18 | 22 | - |
| **C1 lemma IncSc** | 19 | > 24 | 5 |
| 3SG pronoun IncSc | 20 | > 24 | - |
| **N to V** | 21 | > 24 | 20 |
| OOV IncSc | - | 9 | - |
| LIX | - | 12 | - |
| Extra-long token | - | 13 | 6 |
| Lex T to Nr T | - | 14 | 15 |
| PR to PP | - | 16 | - |
| Past V to V | - | 17 | 19 |
| B1 lemma IncSc | - | 19 | 3 |
| Function W IncSc | - | 20 | - |
| Right DepArc Ratio | - | 21 | - |
| Avg token length | - | 23 | 7 |
| B2 lemma IncSc | - | > 24 | 11 |
| N senses per N | - | > 24 | 21 |
| PR to N | - | > 24 | 22 |
| Nominal ratio | - | > 24 | 23 |
| N IncSc | - | > 24 | 24 |

Table 5: K-best features and their rank across different datasets.

## 6 Conclusion and future work

In this work, we described the results of a feature selection method applied to different language learning related datasets. We found a small number of features that proved useful across all datasets regardless of the length of the linguistic input or the type of relevant language learning skill. We showed that besides lexical frequency and variation, the length of dependencies and the amount and type of verbs carry valuable information for predicting proficiency levels. To our knowledge, the usefulness of single features across receptive and productive L2 data of different sizes has not been previously explored. We aimed at finding the optimal number and types of features to use in order to boost performance for these types of predictions. An improved CEFR level classification is especially important for its integration into NLP applications aiming at on-the-fly assessment of texts or exercise generation. In the future, extending this investigation of feature importances to datasets in other languages could contribute to a deeper understanding about which indicators are more universally useful. Furthermore, the selected subset of features could be evaluated also with the help of teaching experts to confirm their usefulness.

## Acknowledgements

We would like to thank the Royal Swedish Academy of Letters, History and Antiquities for providing a travel grant to the first author via the Wallenberg Foundation.

## References

David Alfter, Yuri Bizzoni, Anders Agebjörn, Elena Volodina, and Ildikó Pilán. 2016. From distributions to labels: A lexical proficiency analysis using learner corpora. Number 130 in Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition, pages 1–7. Linköping University Electronic Press.

Jessie Saraza Barrot. 2015. Comparing the linguistic complexity in receptive and productive modes. *GEMA Online® Journal of Language Studies*, 15(2).

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

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.

António Branco, João Rodrigues, Francisco Costa, João Silva, and Rui Vaz. 2014. Rolling out text categorization for language learning assessment supported by language technology. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language*, pages 256–261. Springer.

Kyle Carbon, Kacyn Fujii, and Prasanth Veerina. 2014. Applications of machine learning to predict Yelp ratings. Stanford Univ., Stanford, CA.

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

Scott A Crossley and Danielle S McNamara. 2011. Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life Long Learning*, 21(2-3):170–191.

Pedro Curto, Nuno J Mamede, and Jorge Baptista. 2015. Automatic text difficulty classifier – Assisting the selection of adequate reading materials for European Portuguese teaching. In *Proceedings of the International Conference on Computer Supported Education*, pages 36–44.

Cecilia Fasth and Anita Kannermark. 1997. *Form i focus: övningsbok i svensk grammatik. Del B*. Folkuniv. Förlag, Lund.

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.

Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: a CEFR-graded lexical resource for Swedish foreign and second language learners. In *Proceedings of the 10$^{th}$ International Conference on Language Resources and Evaluation*.

Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods*, 36(2):193–202.

Julia Hancke and Detmar Meurers. 2013. Exploring CEFR classification for German based on rich linguistic modeling. In *Learner Corpus Research Conference*, pages 54–56.

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 the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 460–467.

Katarina Heimann Mühlenbock. 2013. I see what you mean—assessing readability for specific target groups. *Data linguistica*, (24).

Yi-Ting Huang, Hsiao-Pei Chang, Yeali Sun, and Meng Chang Chen. 2011. A robust estimation scheme of reading difficulty for second language learners. In *11$^{th}$ IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 58–62. IEEE.

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

Nikolay Karpov, Julia Baranova, and Fedor Vitugin. 2014. Single-sentence readability prediction in Russian. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 91–100. Springer.

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. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Ildikó Pilán, Sowmya Vajjala, and Elena Volodina. 2016a. A readable read: automatic assessment of language learning materials based on linguistic complexity. *International Journal of Computational Linguistics and Applications (IJCLA)*, 7(1):143–159.

Ildikó Pilán, Elena Volodina, and Lars Borin. 2016b. Candidate sentence selection for language learning exercises: from a comprehensive framework to an empirical evaluation. *Traitement Automatique des Langues (TAL) Journal, Special issue on NLP for Learning and Teaching*, 57(3):67–91.

Ildikó Pilán, Elena Volodina, and Torsten Zesch. 2016c. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of the $26^{th}$ International Conference on Computational Linguistics*, pages 2101–2111.

Robert Reynolds. 2016. Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. In *Proceedings of the $11^{th}$ Workshop on Innovative Use of NLP for Building Educational Applications*, pages 289–300.

Elizabeth Salesky and Wade Shen. 2014. Exploiting morphological, grammatical, and semantic correlates for improved text difficulty assessment. In *Proceedings of the $9^{th}$ Workshop on Innovative Use of NLP for Building Educational Applications*, pages 155–162, June.

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

Yao-Ting Sung, Wei-Chun Lin, Scott Benjamin Dyson, Kuo-En Chang, and Yu-Chia Chen. 2015. Leveling L2 texts through readability: combining multilevel linguistic features with the CEFR. *The Modern Language Journal*, 99(2):371–391.

Anaïs Tack, Thomas François, Sophie Roekhaut, and Cédrick Fairon. 2017. Human and automated CEFR-based grading of short answers. In *Proceedings of the $12^{th}$ Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179.

Sowmya Vajjala and Kaidi Lõo. 2014. Automatic CEFR level prediction for Estonian learner text. *NEALT Proceedings Series Vol. 22*, pages 113–127.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the $7^{th}$ Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–173.

Elena Volodina and Sofie Johansson Kokkinakis. 2012. Introducing the Swedish Kelly-list, a new lexical e-resource for Swedish. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 1040–1046.

Elena Volodina, Ildikó Pilán, Stian Rødven Eide, and Hannes Heidarsson. 2014. You get what you annotate: a pedagogically annotated corpus of coursebooks for Swedish as a second language. In *Proceedings of the $3^{rd}$ workshop on NLP for Computer Assisted Language Learning*, pages 128–144.

Elena Volodina, Ildikó Pilán, Lorena Llozhi, Baptiste Degryse, and Thomas François. 2016a. SweLLex: second language learners' productive vocabulary. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, number 130, pages 76–84. Linköping University Electronic Press.

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016b. SweLL on the rise: Swedish learner language corpus for European Reference Level studies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*.

Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. 2011. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the $11^{th}$ Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22.

Lixiao Zhang, Zaiying Liu, and Jun Ni. 2013. Feature-based assessment of text readability. In *$7^{th}$ International Conference on Internet Computing for Engineering and Science (ICICSE)*, pages 51–54. IEEE.