Assessing the Readability of Sentences: Which Corpora and Features?

Abstract

The paper focuses on sentence readability assessment modelled as a classification task, with the final aim of shedding light on the type of corpora to be used for training and the typology of multi–level linguistic features underlying it.

1 Introduction

Over the last ten years, work on automatic readability assessment employed sophisticated NLP techniques (such as syntactic parsing and statistical language modeling) to capture highly complex linguistic features, and used statistical machine learning to build readability assessment tools. A variety of different NLP–based approaches has been proposed so far in the literature, differing at the level of the number of identified readability classes, the typology of features taken into account, the intended audience of the texts under evaluation, or the application within which readability assessment is carried out, etc.

As pointed out by Skory and Eskenazi (2010), so far most research focused on automatic readability assessment at the document level: methods developed perform well when the task is characterizing the readability level of an entire document, while they are unreliable for short texts, including single sentences (Dell’Orletta et al., 2011; Sjöholm, 2012). Yet, for specific applications, assessing the readability level of individual sentences would be desirable. This is the case, for instance, for text simplification: in the current approaches, text readability is typically assessed with respect to the entire document, while text simplification is carried out at the sentence level, as e.g., done in Aluísio et al. (2010), Bott and Saggion (2011) and Inui et al. (2003). However, by decoupling the readability assessment and simplification processes, the impact of simplification operations on the overall readability level of a given text may not always be clear. With sentence–based readability assessment, this is expected to be no longer a problem. Sentence readability assessment thus represents, in our opinion, an open issue in the literature which is worth being further explored.

In this paper, we focus on sentence readability assessment modelled as a classification task, with the final aim of shedding light on two open issues connected with it, namely the type of corpora to be used for training and the typology of features underlying it. For what concerns the former, sentence readability assessment poses the remarkable issue of classifying sentences according to their difficulty: this fact has important implications at the level of the composition of the corpora to be used for training. The second issue is concerned with whether and to what extent the features playing a significant role in the assessment of readability at the sentence level coincide with those exploited at the level of document. In particular, the following research questions will be tackled: a) in assessing sentence readability, is it better to use a small gold standard training corpus of manually classified sentences or a much bigger training corpus automatically constructed from readability–tagged documents possibly containing misclassified sentences? b) which are the features maximizing sentence readability assessment? c) to what extent do important features for sentence readability classification match those playing a role in the document readability classification?

We will try to answer these questions by working on Italian, which is a less–resourced language as far as readability is concerned, with a version of READ–IT (Dell’Orletta et al., 2011) which we adapted by integrating a specialized training corpus and a maximum entropy–based feature–selection and ranking algorithm (i.e. grafting, discussed below). READ–IT represents the first NLP–based readability assessment tool for Italian,
which has already been released in different versions to comply with the specific needs of different applicative scenarios (Dell’Orletta et al., 2014)

The paper is organized as follows: Section 2 describes the background literature on the topic, Section 3 introduces our approach to the sentence-based readability assessment task, and Section 4 illustrates the exploited linguistic features. Finally, Sections 7 and 8 describe the experimental setting and discuss our results.

2 Background

In spite of the need of performing readability assessment at the sentence level (Skory and Eskenazi, 2010), it appears that adequate sentence-level readability assessment metrics are still lacking. To our knowledge, very few attempts have been made to systematically investigate the main issues and challenges concerned with the readability assessment of sentences (as opposed to documents). For example, we know of only two studies focusing on languages other than English, namely Italian (Dell’Orletta et al., 2011) and Swedish (Sjöholm, 2012). In both cases, the authors start from the assumption that while all sentences occurring in simplified texts can be assumed to be easy-to-read sentences, the reverse is not true, since not all sentences occurring in complex texts are difficult-to-read. This has important consequences at the level of the evaluation of sentence classification results: i.e. erroneous readability assessments within the class of difficult-to-read texts may either correspond to those easy-to-read sentences occurring within complex texts or represent real classification errors. To overcome this problem in the readability assessment of individual sentences, a notion of distance with respect to easy-to-read sentences was introduced by Dell’Orletta et al. (2011).

By contrast, different types of sentence-based analyses are reported in the literature. For instance in a text simplification scenario by Drndarević et al. (2013), Aluísio et al. (2008), Štajner and Saggion (2013) and ERNESTA (Enhanced Readability through a Novel Event-based Simplification Tool, Barlacchi and Tonelli (2013)) which represents the first sentence simplification system released for Italian. Louis and Nenkova (2013) focussed on sentence analysis to predict writing quality level. Sheikha and Inkpen (2012) report the results of both document- and sentence-based classification in the different, but related task of assessing formal vs. informal style of a document/sentence. For students learning English, Andersen et al. (2013) made a self-assessment and tutoring system available which was able to assign a quality score for each individual sentence they write: this provides automated feedback on learners’ writing.

A further important issue, largely investigated in previous readability assessment studies, is the selection of linguistic factors playing a role in assessing the readability of documents. If traditional readability metrics (see e.g., Kincaid et al. (1975)) typically rely on raw text characteristics, such as word and sentence length, the new NLP-based readability indices exploits wider sets of features ranging across different linguistic levels. Starting from Schwarm and Ostendorf (2005) and Heilman et al. (2007), the role of syntactic features in this task was considered, and more recently, the role of discourse features (e.g., discourse topic, discourse cohesion and coherence) has been taken into account (see e.g., Barzilay and Lapata (2008), Pitler and Nenkova (2008), Kate et al. (2010), Feng et al. (2010) and Tonelli et al. (2012)). In many of these studies the importance of each linguistic level has been investigated as well. For example, Feng et al. (2010) systematically evaluated a wide range of features and compared the results of different statistical classifiers trained on different classes of features. Similarly, the correlation between linguistic features captured at different levels of linguistic analysis has been calculated by Pitler and Nenkova (2008) with respect to human readability judgments, and by François and Fairon (2012) with respect to readability levels. In both cases, the classes of features which were highly correlated with readability judgments have been used in a readability assessment task to test their efficacy. However, in all studies, the predictive power of the selected features was evaluated at the document level.

3 Overview of our Approach

In this paper, we tackle the challenge of assessing the readability of individual sentences as a first step towards text simplification. In particular, in this paper we focus on two open issues: the need for reference corpora to be used for training (i.e. collections of sentences classified according to their readability level), and the identification of
the most effective features to determine sentence readability.

For our training corpus, we started from the assumption that an easy-to-read newspaper such as Due Parole (henceforth, 2Par), which includes articles specifically written by Italian linguists experts in text simplification for an audience of adults with a rudimentary literacy level or with mild intellectual disabilities (Piemontese, 1996), should only contain easy-to-read sentences. In contrast, a rather difficult newspaper such as La Repubblica (henceforth, Rep) does not only contain difficult-to-read sentences. Consequently, whereas 2Par is homogeneous at the sentence level, this is not the case for Rep. To assess the impact of the noise in the training set, we constructed different training sets in this study differing in size and internal composition (e.g., a noisy set which assumes all Rep sentences to be difficult-to-read, or a clean, but smaller set in which the easy-to-read sentences where filtered out) and we evaluated the performance in a sentence-based readability assessment task (see Section 7.2).

To find the most predictive features in sentence readability assessment, we used GRAFTING (Perkins et al., 2003) as our incremental feature-selection method (described in Section 6). In this framework, incremental feature selection is added to the training of a maximum entropy model. By comparing the results on the basis of sentence-vs. document-based readability assessment we are able to highlight which and how many features are needed for the assessment of sentence readability and how these correspond to the features needed for document readability.

4 Linguistic Features

The features used by READ-IT for predicting readability are organised into four main categories: raw text features, lexical features, morpho-syntactic features, and syntactic features. This proposed four-fold partitioning closely follows the different levels of linguistic analysis automatically carried out on the text being evaluated, i.e. tokenization, lemmatization, part-of-speech (POS) tagging and dependency parsing. Such a partitioning thus was meant to identify those easy-to-extract features with high discriminative power in order to reduce the linguistic pre-processing of texts, while simultaneously guaranteeing a reliable readability assessment.

Raw Text Features

The raw text features (Features [1–2] in Table 4) refer to those features typically used within traditional readability metrics. They include sentence length, calculated as the average number of words per sentence, and word length, calculated as the average number of characters per words.

Lexical Features

Basic Italian Vocabulary rate features (Features [3–6] in Table 4): these features refer to the internal composition of the vocabulary of the text. To this end, we took as a reference resource the Basic Italian Vocabulary by De Mauro (2000), including a list of 7000 words highly familiar to native speakers of Italian. In particular, we calculated two different features corresponding to: i) the percentage of all unique words (types) on this reference list (calculated on a per-lemma basis), and ii) the internal distribution of the occurring basic Italian vocabulary words into the usage classification classes of ‘fundamental words’ (very frequent words), ‘high usage words’ (frequent words) and ‘high availability words’ (relatively lower frequency words referring to everyday objects or actions and thus well known to speakers). Whereas the latter represents a novel feature in the readability assessment literature, the former originates from the Dale-Chall formula (Chall and Dale, 1995) and, as implemented here, can be seen as the complement of the type out-of-vocabulary rate features used by Petersen and Ostendorf (2009) or by François and Fairon (2012).

Type/Token Ratio (Features [7–8] in Table 4): as stated in Bowers (2000), word repetition may affect the readability of a text. This dimension of variation of a text is typically monitored by computing the Type/Token Ratio (TTR), referring to the ratio between the number of lexical types and the number of tokens within a text. This feature has already been used for readability assessment purposes by e.g., (Aluísio et al., 2010) and (François and Fairon, 2012). Due to its sensitivity to sample size, this feature is computed for text samples of equivalent length.

Morpho-syntactic features

Language model probability of POS unigrams (Features [9–22] in Table 4): this feature is based on a unigram language model assuming that the probability of a token is independent of its context. The model is simply defined by a list of POS types
and their individual probabilities. This feature has already been shown to be a reliable indicator for automatic readability assessment (see, e.g., Pitler and Nenkova (2008) and Aluísio et al. (2010)).

**Lexical density** (Feature [23] in Table 4): this feature refers to the ratio of content words (verbs, nouns, adjectives and adverbs) to the total number of lexical tokens in a text. Content words have been used for readability assessment by Aluísio et al. (2010) and Feng et al. (2010).

**Verbal mood** (Features [24–41] in Table 4): this feature refers to the distribution of verbs by mood and/or tense. Following Carreiras et al. (1997), this information type appears to play an important role in the construction of mental models while reading. It is a language-specific feature exploiting the predictive power of the rich verbal morphology, as demonstrated by Dell’Orelta et al. (2011) for Italian and François and Fairon (2012) for French.

**Syntactic Features**

**Unconditional probability of dependency types** (Features [42–70] in Table 4): this feature refers to the unconditional probability of different types of syntactic dependencies (e.g., subject, direct object, modifier, etc.) and can be seen as the dependency-based counterpart of the ‘phrase type rate’ feature used by Nenkova et al. (2010).

**Parse tree depth features** (Features [71–72] in Table 4): parse tree depth can be indicative of increased sentence complexity as stated by, e.g., Yngve (1960), Frazier (1985) and Gibson (1998). This set of features is meant to capture different aspects of the parse tree depth and includes the following measures: a) the depth of the whole parse tree, calculated in terms of the longest path from the root of the dependency tree to some leaf, and b) the average depth of embedded complement ‘chains’ governed by a nominal head and including either prepositional complements or nominal and adjectival modifiers. The first feature has already been used in syntax-based readability assessment studies (Schwarm and Ostendorf, 2005; Heilman et al., 2007; Nenkova et al., 2010), and the latter is reminiscent of the ‘head noun modifiers’ feature used by Nenkova et al. (2010).

**Verbal predicate features** (Features [73–78] in Table 4): this set of features captures different aspects of the behaviour of verbal predicates. They range from the number of verbal roots with respect to number of all sentence roots occurring in a text to their arity, meant as the number of instantiated dependency links sharing the same verbal head (covering both arguments and modifiers). The relevance of this feature for readability assessment was demonstrated by Kintsch et al. (1975) who showed that the number of propositions as well as the number of different arguments in a sentence influence its reading time. To our knowledge, READ-IT (Dell’Orelta et al., 2011) was the first system using this feature for readability assessment purposes, followed, more recently, by François and Fairon (2012) and Falkenjack et al. (2013). Within this feature set, we also consider the relative ordering of subject and object with respect to the verbal head. Even if Italian is a free word order language, the subject–verb–object order is preferred, as also empirically demonstrated by Dell’Orelta et al. (2005). Even though word order is not an effective cue for subject/object identification in Italian, sentences characterized by a non–canonical order (i.e. with subject in post–verbal rather than pre–verbal position or with pre–verbal objects rather than post–verbal ones) are more difficult to read.

**Subordination features** (Features [79–85] in Table 4): subordination is widely acknowledged to be an index of structural complexity in language as showed in Aluísio et al. (2010) and Dell’Orelta et al. (2011). A first feature was meant to measure the distribution of subordinate vs. main clauses. For subordinates, we also considered the distribution of infinitives vs finite complement clauses and their relative ordering with respect to the main clause. According to Miller and Weinert (1998), sentences containing subordinate clauses in post–verbal rather than in pre–verbal position are easier to read. Since it is a widely acknowledged fact that highly complex sentences contain deeply embedded subordinate clauses, the average depth of ‘chains’ of embedded subordinate clauses was also computed.

**Length of dependency links feature** (Features [86–87] in Table 4): both Lin (1996) and Gibson (1998) showed that the syntactic complexity of sentences can be predicted with measures based on the length of dependency links. This is also demonstrated in McDonald and Nivre (2007) who claim that statistical parsers have a drop in accuracy when analysing long–range dependencies. Here, the dependency length is measured in terms of the words occurring between the syntactic head
and the dependent. This feature is the dependency-based counterpart of the ‘phrase length’ feature used for readability assessment by Nenkova et al. (2010) and Feng et al. (2010).

5 Corpora

In our experiments, we relied on two different corpora: a newspaper corpus, La Repubblica (Rep), and an easy-to-read newspaper, Due Parole (2Par). As discussed by Dell’Orletta et al. (2011), the two corpora – selected as representative of complex vs. simplified texts within the journalistic genre – differ significantly with respect to the distribution of features typically correlated with text complexity.

However, as pointed out in Section 3, whereas such a distinction is valid as far as documents are concerned, it does not need to be valid with respect to sentences. In other words, whereas we can consider all sentences of 2Par as easy-to-read, not all Rep sentences have to be difficult-to-read. To overcome this asymmetry, we constructed a corpus of difficult-to-read sentences by manually annotating the Rep sentences according to their readability (easy vs. difficult). The annotation process was carried out by two annotators with a background in computational linguistics. In order to assess the reliability of their judgements, we started with a small annotation experiment: the two annotators were provided with the same 5 articles from the Rep corpus (for a total of 107 sentences) and were asked to extract the difficult-to-read sentences (as opposed to both easy-to-read and not-easy-to-classify sentences). The first annotator carried out the task in 5 minutes and 46 seconds, while the second annotator took 9 minutes and 8 seconds. The two annotators agreed on the classification of 81 difficult-to-read sentences out of 107 considered ones (in particular, the first annotator identified 90 difficult-to-read sentences and the second one 93 sentences), corresponding to about 90% of sentences classified as difficult-to-read. Given this high agreement, the two annotators were asked to select difficult sentences from two sets of distinct articles. This resulted in a combined set of 1,745 difficult-to-read sentences\(^1\), which were used (together with a random selection of easy-to-read sentences from 2Par) for training and testing (see Section 7).

6 Model Training and Feature Ranking

Given the twofold goal of this study, i.e. reliably assessing sentence readability and identifying the most relevant features contributing to it, READ-IT was adapted to meet both goals. Within READ-IT GRAFTING (Perkins et al., 2003) was used to train a maximum entropy model while simultaneously including incremental feature selection. This method uses a gradient-based heuristic to select the most promising feature (to add to the set of selected features $S$), and then performs a full weight optimization over all features in $S$. This process is repeated until a certain stopping criterion is reached. As the grafting approach we use integrates the $l_1$ regularization (preventing overfitting), features are only included (i.e. have a non-zero weight) when the reduction of the objective function is greater than a certain threshold. In our case, the $l_1$ prior we use was selected on the basis of evaluating maximum entropy models with varying $l_1$ values (range: 1e-11, 1e-10, ..., 0.1, 1) via 10-fold cross validation.

We used TINYEST\(^2\), a grafting-capable maximum entropy parameter estimator for ranking tasks (de Kok, 2011; de Kok, 2013), to select the features and estimate their weights. Whereas our task is not a ranking task, but rather a binary classification problem, we were able to model it as a ranking task by assigning a high score (1) to difficult-to-read sentences and a low score (0) to easy-to-read sentences. Consequently, a sentence having a score $< 0.5$ was interpreted as an easy-to-read sentence, whereas a sentence which was assigned a score $\geq 0.5$ was interpreted to be a difficult-to-read sentence.

7 Experiments and Results

7.1 Experimental Setup

In all experiments, the corpora were automatically POS tagged by the part-of-speech tagger described in Dell’Orletta (2009) and dependency-parsed by the DeSR parser (Attardi, 2006) using Support Vector Machines as learning algorithm.

We devised two different experiments. To answer our first research question, the first experiment (the results of which are described in Section 7.2) aimed at identifying the best training data for sentence readability assessment. Our goal here was to compare the results on the basis of using a

\(^1\)The collection can be downloaded from www.italianlp.it.

\(^2\)http://github.com/danieldk/tinyest
small set of gold standard data with respect to a (potentially larger, but) noisy dataset (i.e. without manual revision) where every Rep sentence was assumed to be difficult-to-read. In particular, the comparison involved four datasets:

- a collection of gold standard data consisting of 1,310 easy-to-read sentences randomly extracted from the 2Par corpus and 1,310 manually selected difficult-to-read sentences from the Rep corpus;
- a large collection of uncorrected data consisting of the whole 2Par corpus (3,910 easy-to-read sentences) and the whole Rep corpus (8,452 sentences, classified a priori as difficult-to-read);
- a balanced collection of uncorrected sentences, consisting of 3,910 sentences from 2Par and 3,910 sentences from Rep;
- a balanced collection of uncorrected sentences with the same size as the gold standard dataset, namely 1,310 sentences from 2Par and 1,310 sentences from Rep.

The four models we built were evaluated using a held-out test set consisting of 435 sentences from 2Par and 435 manually classified difficult-to-read sentences from the Rep corpus. Using the grafting method explained in Section 6, we calculated the classification score for each sentence in our test set on the basis of an increasing number of features (ranging from 1 to all non-zero weighted features for the specific dataset): sentences with a score below 0.5 were classified as easy-to-read, whereas

Table 1: Ranking position of the features in sentence and document readability assessment experiments. Only about 14 features are needed for an adequate model of document readability, whereas this number increases to 30 for sentence readability (marked in boldface). Features which were not selected during our ranking approach have no rank.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Sent. class.</th>
<th>Doc. class.</th>
<th>Feature</th>
<th>Sent. class.</th>
<th>Doc. class.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Sentence length</td>
<td>1</td>
<td>1</td>
<td>2) Word length</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

**Lexical features:**

- 3) Word types in the BNC Italian Vocabulary | 14 | 24 |
- 4) Fundamental words | 10 | 9 |
- 5) High usage words | 22 | 38 |

**Morpho-syntactic features:**

- 6) Main verb – cond. mood | 44 | 37 |
- 7) Main verb – part. mood | 24 | 8 |
- 8) Main verb – subj. mood | 48 | 36 |
- 9) Main verb – imp. mood | 54 | 29 |
- 10) Main verb – ger. mood | 56 | 25 |
- 11) Main verb – inf. mood | 67 | 57 |

**Syntactic features:**

- 12) Argument | 62 | 65 |
- 13) Auxiliary | 70 | 66 |
- 14) Clitic | 63 | 57 |
- 15) Preposition | 20 | 20 |
- 16) Locative complement | 29 | 29 |
- 17) Relative modifier | 31 | 31 |

**Raw text features:**

- 18) Sentence root | 35 | 35 |
- 19) Subordinate clause | 39 | 52 |
- 20) Temporal predicate | 64 | 64 |
- 21) Embedded complement/chauds | 66 | 66 |
- 22) Embedded complement/chauds | 68 | 68 |
- 23) Pre-verbal subject | 70 | 70 |
- 24) Pre-verbal object | 72 | 72 |
- 25) Main clauses | 75 | 75 |

**Sent. class.**

- 26) Auxiliary word | 61 | 61 |
- 27) Aux. verb – subj. mood | 47 | 47 |
- 28) Aux. verb – cond. mood | 20 | 20 |
- 29) Aux. verb – part. mood | 36 | 36 |
- 30) Subordinate clause | 59 | 59 |
- 31) Subordinate clause | 60 | 60 |
- 32) Complex | 33 | 33 |
- 33) Complex | 34 | 34 |
- 34) Complex | 35 | 35 |
- 35) Complex | 36 | 36 |
- 36) Complex | 37 | 37 |
- 37) Complex | 38 | 38 |
- 38) Complex | 39 | 39 |
- 39) Complex | 40 | 40 |
- 40) Complex | 41 | 41 |
- 41) Complex | 42 | 42 |
- 42) Complex | 43 | 43 |
- 43) Complex | 44 | 44 |
- 44) Complex | 45 | 45 |
- 45) Complex | 46 | 46 |
- 46) Complex | 47 | 47 |
- 47) Complex | 48 | 48 |
- 48) Complex | 49 | 49 |
- 49) Complex | 50 | 50 |
- 50) Complex | 51 | 51 |
- 51) Complex | 52 | 52 |
- 52) Complex | 53 | 53 |
- 53) Complex | 54 | 54 |
- 54) Complex | 55 | 55 |
- 55) Complex | 56 | 56 |
- 56) Complex | 57 | 57 |
- 57) Complex | 58 | 58 |
- 58) Complex | 59 | 59 |
- 59) Complex | 60 | 60 |
- 60) Complex | 61 | 61 |
- 61) Complex | 62 | 62 |
- 62) Complex | 63 | 63 |
- 63) Complex | 64 | 64 |
- 64) Complex | 65 | 65 |
- 65) Complex | 66 | 66 |
- 66) Complex | 67 | 67 |
- 67) Complex | 68 | 68 |
- 68) Complex | 69 | 69 |
- 69) Complex | 70 | 70 |
- 70) Complex | 71 | 71 |
- 71) Complex | 72 | 72 |
- 72) Complex | 73 | 73 |
- 73) Complex | 74 | 74 |
- 74) Complex | 75 | 75 |
- 75) Complex | 76 | 76 |
- 76) Complex | 77 | 77 |
- 77) Complex | 78 | 78 |
- 78) Complex | 79 | 79 |
- 79) Complex | 80 | 80 |
- 80) Complex | 81 | 81 |
- 81) Complex | 82 | 82 |
- 82) Complex | 83 | 83 |
- 83) Complex | 84 | 84 |
- 84) Complex | 85 | 85 |
- 85) Complex | 86 | 86 |
- 86) Complex | 87 | 87 |

**Doc. class.**

- 87) Maximum length of dependency links | 7 | 7 |
sentences having a score greater or equal to 0.5 were classified as difficult–to–read. This procedure was repeated for each of the four models.

The second experiment (results of which are reported in Section 7.3) was aimed at answering our second and third research questions, focusing on the features important for sentence readability, and the relationship of those features with document readability classification. For this purpose, we compared sentence– and the document–based readability classification results. In particular, we compared the features used by the sentence–based readability model trained on the gold standard data and the features used by the document–based model trained on Rep and 2Par. With respect to the document classification, we used a corpus of 638 documents (319 extracted from 2Par representing easy–to–read texts, and 319 extracted from Rep representing difficult–to–read texts) with 20% of the documents constituting the test set.

### 7.2 Training Corpus for Sentence Classification

Table 2 reports the READ–IT results for the sentence classification task using the four training datasets described above. All training models were tested using an increasing number of features (10, 30, 50 and all features) as resulting from the GRAFTING–based ranking. Clearly, the classification model trained on the small gold standard dataset generally outperforms all other models: it achieves the best accuracy (83.7%) using a relatively small number of features (66), and also for a fixed number of features (30 and 50). Only when using the top–10 features, the uncorrected balanced big dataset slightly outperforms the gold standard dataset.

Of course, it is clear that the improvements of using the gold standard dataset over the uncorrected training datasets are modest. Consequently, treating the Rep corpus as a collection of difficult–to–read sentence is not completely unjustified. This is further illustrated by the satisfactory results reported by Dell’Orletta et al. (2011). Nevertheless, in this study we demonstrated that the readability assessment accuracy can be improved by using a manually selected (small) training dataset, and this also suggests that better results can be obtained by constructing a larger gold standard training dataset.

### 7.3 Sentence vs Document Classification: which and how many features?

In this section, we focus on how many and which features are needed in the sentence–based readability assessment task, and compare these to those needed for document–based readability assessment. For this purpose, we compared the results obtained by the grafting–based feature selection in the sentence classification task (using the gold standard dataset for training) to those obtained in the document classification task (here, the gold standard consisted of treating all 2Par texts as easy–to–read and all Rep texts as hard–to–read). Comparable to Table 2, Table 3 shows the accuracy on the test set (in this case, consisting of 319 easy–to–read and 319 hard–to–read documents) for an increasing number of features selected via GRAFTING.

<table>
<thead>
<tr>
<th>Training data</th>
<th>10 ft.</th>
<th>30 ft.</th>
<th>50 ft.</th>
<th>all ft.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unbalanced large</td>
<td>63.7</td>
<td>74.9</td>
<td>78.4</td>
<td>78.9 (85 ft)</td>
</tr>
<tr>
<td>Balanced small</td>
<td>67.9</td>
<td>79.2</td>
<td>80.8</td>
<td>82.5 (82 ft)</td>
</tr>
<tr>
<td>Balanced large</td>
<td>70.6</td>
<td>79.7</td>
<td>81.0</td>
<td>82.3 (85 ft)</td>
</tr>
<tr>
<td>Gold standard</td>
<td>69.8</td>
<td>79.9</td>
<td>81.3</td>
<td>83.7 (66 ft)</td>
</tr>
</tbody>
</table>

Table 2: Accuracy of the sentence classification task using four different training data sets and a varying number of features.

<table>
<thead>
<tr>
<th>Training data</th>
<th>10 ft.</th>
<th>30 ft.</th>
<th>50 ft.</th>
<th>70 ft. (all)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rep and 2Par</td>
<td>93.3</td>
<td>96.6</td>
<td>96.6</td>
<td>95</td>
</tr>
</tbody>
</table>

Table 3: Accuracy of the document classification task for a varying number of features.

Clearly, the best document classification results are obtained using a set of 30 features. In contrast to sentence classification where adding features always increases performance on the test set, more features do not appear to help for document classification. Thus sentence readability classification seems to be more complex, requiring a higher amount of features. This trend is also highlighted in Figures 1 and 2, where we visualize the classification results on the training set (using 10–fold cross–validation) and the held–out test set for an increasing amount of features. As Figure 1 shows, the document classification task requires about 14 features after which the performance appears to stabilize. In contrast, Figure 2 shows that sentence classification requires at least 30 features.

There are clear differences in the ranked list of features. Table 4 reports the comparison be-
between the ranking of the features obtained in the sentence– and document–based classification experiments. Interestingly, in both experiments the raw text features *Sentence length* and *Word length* are the top-two features. However, in contrast to the findings of Dell’Orletta et al. (2011), the grafting process showed that both in sentence– and document–based readability assessment, syntactic features play a central role. Most of these are highly ranked, with some striking differences. For example, *Chains* of embedded subordinate clauses [83], *Subordinate clauses in a pre–verbal position* [81], and the *Length of dependency links* [86] turned out to be mainly relevant for document readability. In contrast, syntactic features such as *Arity of verbal predicates* [74], *Pre–verbal subject* [75], *Post–verbal object* [78], and *Embedded complement ‘chains’* [72] are mainly useful for sentence readability. Some play a similar role in both readability classification tasks. This is the case, for example, for *Verbal root* [73], *Parte tree depth* [71], and *Max. length of dependency links* [87]. Interestingly, at the morpho–syntactic level, the feature ranking is much more comparable, as the numbers are much closer between the two columns. Lexical features show a much more mixed result. Type-Token Ratio is only important for document classification, whereas most of the other features are important for sentence readability, but not for document readability (with the exception of the presence of “fundamental words”).

![Figure 1: Document classification results.](image1)

8 Discussion

In this study we have focused on three research questions. First, we asked which type of corpus is best to assess sentence readability. Whereas we found that using a set of manually selected complex sentences was better than using a simple corpus–based distinction, the extra effort needed to construct the training corpus might not be worthwhile as improvements were only modest. However, we did not consider at a more sensitive measure of the difficulty of a sentence (such as a number ranging between 0 and 1), and this might be able to offer a more substantial improvement (at the cost of needing more time to create the training material). Of course, when the goal is to identify the best features for assessing sentence readability, it does make sense to have high-quality training data to prevent selecting inadequate features.

The second research question involved identifying which features were most useful for assessing sentence readability. Besides raw text features, syntactic but also morpho–syntactic features turned out to play a central role to achieve adequate performance.

The third research question investigated the overlap between the features needed for document and sentence readability classification. Whereas there certainly was overlap between the top features, most of the features had a different rank across the two tasks, and specifically suggested that the sentence readability task is more complex than that of document readability. Of course this makes sense, given that there is much less information available for a sentence than for a document (containing a lot of sentences).

![Figure 2: Sentence classification results.](image2)

References


Fadi Abu Sheikha and Diana Inkpen. 2012. Learning to classify documents according to formal and informal style. volume 8.


