# Assessing the Readability of Sentences: Which Corpora and Features?

**Felice Dell'Orletta**[◇], **Martijn Wieling**[•◦], **Andrea Cimino**[◇], **Giulia Venturi**[◇]
**and Simonetta Montemagni**[◇]
[◇]Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{name.surname}@ilc.cnr.it
[•]Department of Humanities Computing, University of Groningen, The Netherlands
[◦]Department of Quantitative Linguistics, University of Tübingen, Germany
wieling@gmail.com

## Abstract

The paper investigates the problem of sentence readability assessment, which is modelled as a classification task, with a specific view to text simplification. In particular, it addresses two open issues connected with it, i.e. the corpora to be used for training, and the identification of the most effective features to determine sentence readability. An existing readability assessment tool developed for Italian was specialized at the level of training corpus and learning algorithm. A maximum entropy–based feature selection and ranking algorithm (grafting) was used to identify to the most relevant features: it turned out that assessing the readability of sentences is a complex task, requiring a high number of features, mainly syntactic ones.

## 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.

Research focused so far on readability assessment at the document level. However, as pointed out by Skory and Eskenazi (2010), 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. Yet, for specific applications, assessing the readability level of individual sentences would be desirable. This is the case, for instance, for text simplification: in 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). 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 an open issue in the literature which is worth being further explored. To our knowledge, the only attempts in this direction are represented by Dell'Orletta et al. (2011) and Sjöholm (2012) for the Italian and Swedish languages respectively, followed more recently by Vajjala and Meurers (2014) dealing with English.

In this paper, we tackle the challenge of assessing the readability of individual sentences as a first step towards text simplification. The task is modelled as a classification task, with the final aim of shedding light on two open issues connected with it, namely the 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 what concerns the former, sentence readability assessment poses the remarkable issue of classifying sentences according to their difficulty: if all sentences occurring in simplified texts can be assumed to be easy–to–read sentences, the reverse does not necessarily hold since not all sentences occurring in complex texts are to be assumed difficult–to–read. This fact has important implications at the level of the composition of the corpora to be used for training. The sec-



*Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications* , pages 163–173,
Baltimore, Maryland USA, June 26, 2014. 

ond 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 are addressed:

1. 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?

2. which are the features maximizing sentence readability assessment?

3. 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. To this end, READ–IT (Dell'Orletta et al., 2011; Dell'Orletta et al., 2014), which represents the first NLP–based readability assessment tool for Italian, was specialized in different respects, namely at the level of the training corpus and of the learning algorithm; to investigate questions 2. and 3. above, a maximum entropy–based feature selection and ranking algorithm (i.e. grafting) was selected. The specific target audience of readers addressed in this study is represented by people characterised by low literacy skills and/or by mild cognitive impairment. The paper is organized as follows: Section 2 describes the background literature, Section 3 introduces our approach to the task, in terms of used corpora, features and learning algorithm. Finally, Sections 4 and 5 describe the experimental setting and discuss achieved results.

## 2 Background

In spite of the acknowledged need of performing readability assessment at the sentence level, so far very few attempts have been made to systematically investigate the issues and challenges concerned with the readability assessment of sentences (as opposed to documents). The first two studies in this direction focused 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). Focusing on English, a similar issue is addressed more recently by Vajjala and Meurers (2014) who developed a binary sentence classifier trained on Wikipedia and Simple English Wikipedia: they showed that the low accuracy obtained by their classifier stems from the incorrect assumption that all Wikipedia sentences are more complex than the Simple Wikipedia ones.

Besides readability, sentence–based analyses are reported in the literature for related tasks: for instance, in a text simplification scenario by Drndarević et al. (2013), Aluísio et al. (2008), Štajner and Saggion (2013) and Barlacchi and Tonelli (2013); or to predict writing quality level by Louis and Nenkova (2013). 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 identification 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 exploit 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 also 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)). Many of these studies also explored the usefulness of features belonging to individual levels of linguistic description in predicting text readability. 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 level–specific features 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 turned out to be highly correlated with readability judgments were used in a readability assessment tool to test their efficacy. Note, however, that in all cases the predictive power of the selected features was evaluated at the document level only.

## 3 Our Approach

In this section, we introduce the main ingredients of our approach to sentence readability assessment, corpora used for training and testing, selected features and the learning and feature selection algorithm.

### 3.1 Corpora

We relied on two different corpora: a newspaper corpus, *La Repubblica* (henceforth, *Rep*), and an easy–to–read newspaper, *Due Parole* (henceforth, *2Par*). *2Par* 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), which represents the target audience of this study. 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 (Dell'Orletta et al., 2011) and thus represent reliable training datasets. However, whereas such a distinction is valid as far as documents are concerned, it appears to be a simplistic generalization when the focus is on sentences. In other words, whereas we can consider all sentences of *2Par* as easy–to–read, not all *Rep* sentences are expected to be difficult–to–read. From this it follows that whereas the internal composition of *2Par* is homogeneous at the sentence level, this is not the case for *Rep*.

To overcome this asymmetry and in particular to assess the impact of the noise in the *Rep* training corpus, we constructed different training sets differing in size and internal composition, going from a noisy set which assumes all *Rep* sentences to be difficult–to–read to a clean but smaller set in which the easy–to–read sentences occurring in *Rep* were manually filtered out. These training sets were used in different experiments whose results are reported in Section 4.2.

The corpus containing only difficult–to–read sentences was manually built by annotating *Rep* sentences according to their readability (i.e. 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). The agreement between the two annotators was calculated in terms of precision, by taking one of the annotation sets as the gold standard and the other as response: on average, we obtained a precision of 0.88 in the retrieval of sentences definitely classified as difficult–to–read. Given the high level of agreement, the two annotators were asked to select difficult sentences from two sets of distinct *Rep* articles. This resulted in a set of 1,745 difficult–to–read sentences which were used together with a random selection of easy–to–read sentences from *2Par* for training and testing.[1]

[1]The collection can be downloaded from www.italianlp.it/?page_id=22.

| Feature | Ranking position | | Feature | Ranking position | |
|---|---|---|---|---|---|
| | Sent. class. | Doc. class. | | Sent. class. | Doc. class. |
| **Raw text features:** | | | | | |
| [1] Sentence length | **1** | **1** | [2] Word length | **2** | **2** |
| **Lexical features:** | | | | | |
| [3] Word types in the *Basic Italian Vocabulary* | **14** | 42 | [6] "High availability words" | **21** | 22 |
| [4] "Fundamental words" | **10** | **9** | [7] TTR (form) | | **7** |
| [5] "High usage words" | **22** | 38 | [8] TTR (lemma) | | 53 |
| **Morpho–syntactic features:** | | | | | |
| [9] Adjective | | 46 | [26] Aux. verb – inf. mood | 64 | |
| [10] Adverb | **29** | 59 | [27] Aux. verb – part. mood | 51 | |
| [11] Article | 49 | 25 | [28] Aux. verb – subj. mood | 55 | |
| [12] Conjunction | | 40 | [29] Main verb – cond. mood | 40 | 43 |
| [13] Determiner | 43 | 54 | [30] Main verb – ger. mood | 48 | 48 |
| [14] Interjection | | | [31] Main verb – imp. mood | 37 | 57 |
| [15] Noun | **12** | 19 | [32] Main verb – indic. mood | **16** | **11** |
| [16] Number | 65 | 44 | [33] Main verb – inf. mood | **13** | **13** |
| [17] Predeterminer | | | [34] Main verb – part. mood | **26** | 28 |
| [18] Preposition | 61 | | [35] Main verb – subj. mood | 46 | 32 |
| [19] Pronoun | **27** | 30 | [36] Modal verb - inf. mood | 54 | 56 |
| [20] Punctuation | | 35 | [37] Modal verb – cond. mood | 41 | 36 |
| [21] Residual | | | [38] Modal verb – imp. mood | | |
| [22] Verb | 63 | 34 | [39] Modal verb – indic. mood | **18** | 23 |
| [23] Lexical density | 34 | 33 | [40] Modal verb – part. mood | | |
| [24] Aux. verb – cond. mood | 59 | 60 | [41] Modal verb – subj. mood | 60 | 58 |
| [25] Aux. verb – indic. mood | **17** | 17 | | | |
| **Syntactic features:** | | | | | |
| [42] Argument | 62 | | [65] Sentence root | 35 | 62 |
| [43] Auxiliary | | 70 | [66] Subject | 39 | 52 |
| [44] Clitic | | 63 | [67] Subordinate clause | | 64 |
| [45] Complement | **28** | 29 | [68] Temporal complement | 45 | 55 |
| [46] Concatenation | | 66 | [69] Temporal modifier | | |
| [47] Conjunct in a disjunctive compound | 58 | 67 | [70] Temporal predicate | | |
| [48] Conjunct linked by a copulative conjunction | 38 | 37 | [71] Parse tree depth | **5** | **4** |
| [49] Copulative conjunction | 31 | 39 | [72] Embedded complement 'chains' | **8** | 24 |
| [50] Determiner | 50 | 26 | [73] Verbal Root | **6** | **3** |
| [51] Direct object | 44 | 27 | [74] Arity of verbal predicates | **3** | 15 |
| [52] Disjunctive conjunction | 57 | 68 | [75] Pre–verbal subject | **4** | **12** |
| [53] Indirect complement/object | 66 | | [76] Post–verbal subject | **25** | 16 |
| [54] Locative complement | 52 | 51 | [77] Pre–verbal object | 36 | 41 |
| [55] Locative modifier | | | [78] Post–verbal object | **9** | 21 |
| [56] Locative predicate | | | [79] Main clauses | **23** | **14** |
| [57] Modal verb | | 61 | [80] Subordinate clauses | 42 | 45 |
| [58] Modifier | **20** | 47 | [81] Subordinate clauses in pre–verbal position | 32 | **10** |
| [59] Negative | 56 | 69 | [82] Subordinate clauses in post–verbal position | **19** | 20 |
| [60] Passive subject | | | [83] 'Chains' of embedded subordinate clauses | **11** | **5** |
| [61] Predicative complement | | 49 | [84] Finite complement clauses | **30** | 18 |
| [62] Preposition | | | [85] Infinitive clauses | 53 | 50 |
| [63] Punctuation | **24** | 31 | [86] Length of dependency links | **15** | **8** |
| [64] Relative modifier | 47 | 65 | [87] Maximum length of dependency links | **7** | **6** |

Table 1: Typology of features and ranking position 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 ranking have no rank.

### 3.2 Linguistic Features

The set of features used in the experiments reported in this paper is wide, spanning across different levels of linguistic analysis. They can be broadly classified into four main classes, as reported in Table 1: raw text features, lexical features, morpho–syntactic features and syntactic features, shortly described below.[2]

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

The cover category of lexical features (Features [3–8] in Table 1) includes features referring to

[2]For an exhaustive discussion including the motivations underlying this selection of features, the interested reader is referred to Dell'Orletta et al. (2011, 2014) where these features were successfully used for assessing the readability of Italian texts.

both the internal composition of the vocabulary and the lexical richness of the text. For what concerns the former, the *Basic Italian Vocabulary* by De Mauro (2000) was taken as a reference resource, including a list of 7000 words highly familiar to native speakers of Italian. In particular, we consider: a) the percentage of all unique words (types) on this reference list occurring in the text, and b) 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 life). Lexical richness of texts is monitored by computing the Type/Token Ratio (TTR), which refers to the ratio between the number of lexical types and the number of tokens within a text. Due to its sensitivity to sample size, this feature is computed for text samples of equivalent length.

The set of morpho–syntactic features (Features [9–41] in Table 1) is aimed at capturing different aspects of the linguistic structure affecting in one way or another the readability of a text. They range from the probability distribution of part–of–speech (POS) types, to the lexical density of the text, calculated as the ratio of content words (verbs, nouns, adjectives and adverbs) to the total number of lexical tokens in a text. This class also includes features referring to the distribution of verbs by mood and/or tense, which can be seen as a language–specific feature exploiting the predictive power of the Italian rich morphology.

The set of syntactic features (Features [42–87] in Table 1) captures different aspects of the syntactic structure which are taken as reliable indicators for automatic readability assessment, namely:

- the unconditional probability of syntactic dependency types, e.g. subject, direct object, modifier, etc. (Features 42–70);
- parse tree depth features (71–72), going from the *depth of the whole parse tree*, calculated in terms of the longest path from the root of the dependency tree to some leaf, to a more specific feature referring to the *average depth of embedded complement 'chains'* governed by a nominal head and including either prepositional complements or nominal and adjectival modifiers;
- verbal predicate features (73–78) aimed at capturing 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 more specific features such as the *arity of verbs*, meant as the number of instantiated dependency links sharing the same verbal head (covering both arguments and modifiers) and the *relative ordering of subject and object with respect to the verbal head*;
- as subordination is widely acknowledged to be an index of structural complexity in language, subordination features (79–85) include: the *distribution of subordinate vs. main clauses*; for subordinates, the *distribution of infinitives vs finite complement clauses*, their *relative ordering with respect to the main clause* and the *average depth of 'chains' of embedded subordinate clauses*;
- the length of dependency links is another characteristic connected with the syntactic complexity of sentences. Features 86–87 measure dependency length in terms of the words occurring between the syntactic head and the dependent: they focus on all dependency links vs. maximum dependency links only.

### 3.3 Model Training and Feature Ranking

Given the twofold goal of this study, i.e. reliably assessing sentence readability and finding the most predictive features undelying it, we used GRAFTING (Perkins et al., 2003), as this approach allows to train a maximum entropy model while simultaneously including incremental feature selection. The 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 treshold. 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[3], a

[3] http://github.com/danieldk/tinyest

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.

## 4 Experiments and Results

### 4.1 Experimental Setup

In all experiments, the corpora were automatically 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, aimed at exploring the research questions investigated in this paper. To this end, READ–IT was adapted by integrating a specialized training corpus and a maximum entropy–based feature selection and ranking algorithm (i.e. grafting).

**Experiment 1**

This experiment, investigating the first research question, is aimed at identifying what is the most effective training data for sentence readability assessment. In particular, the goal is to compare the results on the basis of using a small set of gold standard data with respect to a (potentially larger, but) noisy data set (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 and unbalanced 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 having the same size as the gold standard dataset, namely 1,310 sentences from *2Par* and 1,310 sentences from *Rep*.

To assess similarities and differences at the level of the different corpora used for training in this experiment, in Table 2 we report a selection of linguistic features (see Section 3.2) characterizing the four datasets with respect to the whole *2Par* corpus.We can observe that *2Par* differs from all four *Rep* corpora for all reported features, and that the four *Rep* corpora show similar trends. Interestingly, however, the *Rep* Gold corpus is almost always the most distant one from *2Par* (i.e. at the level of sentence length, word length, distribution of adjectives and subjects, average length of dependency links and parse tree depth).

On the basis of the four *Rep* datasets, four models were built which we evaluated using a held–out test set consisting of 435 sentences from *2Par* and 435 manually classified difficult–to–read sentences from *Rep*. Using the grafting method, 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 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.

**Experiment 2**

The second experiment is aimed at answering our second and third research questions, focusing on the features relevant for sentence readability, and the relationship of those features with document readability classification. For this purpose, we compared sentence– and 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* representating easy–to–read texts, and 319 extracted from *Rep* representing difficult–to–read texts) with 20% of the documents constituting the held–out test set.

| Features | *Rep* Unbalan. large | *Rep* Balan. small | *Rep* Balan. large | *Rep* Gold | *2Par* |
|---|---|---|---|---|---|
| Sentence length | 24.98 | 26.03 | 25.26 | 28.14 | 18.66 |
| Word length | 5.14 | 5.24 | 5.14 | 5.28 | 5 |
| "Fundamental words" | 75.05% | 75.08% | 74.83% | 74.99% | 76.38 |
| Adjective | 6.19% | 6.25% | 6.36% | 6.42% | 6.03% |
| Noun | 25.65% | 27.09% | 25.74% | 26.10% | 29.13% |
| Subject | 4.62% | 4.75% | 4.64% | 4.42% | 6% |
| Max. length of dependency links | 9.73 | 10.13 | 9.85 | 10.98 | 7.67 |
| Parse tree depth | 6.18 | 6.57 | 6.30 | 6.83 | 5.2 |

Table 2: Distribution of some linguistic features in *Rep* and *2Par* training data

| | Accuracy | | | | | Precision (all ft) | |
|---|---|---|---|---|---|---|---|
| **Training data** | **2 ft** | **10 ft** | **30 ft** | **50 ft** | **all ft** | **Easy** | **Difficult** |
| Unbalanced large | 50 | 63.7 | 74.9 | 78.4 | 78.9 (85 ft) | 69.2 | ***88.5*** |
| Balanced small | 64 | 67.9 | 79.2 | 80.8 | 82.5 (82 ft) | 82.5 | 82.5 |
| Balanced large | 63.9 | **70.6** | 79.7 | 81.0 | 82.3 (85 ft) | 83.0 | 81.6 |
| Gold data | **65.6** | 69.8 | **79.9** | **81.3** | **83.7 (66 ft)** | ***84.8*** | 82.5 |

Table 3: Sentence classification results using four training datasets and a varying number of features

### 4.2 Which Training Corpus for Sentence Classification?

Table 3 reports the results for the sentence classification task using the four training datasets described above. Results are reported in terms of both overall accuracy (calculated as the proportion of correct answers against all answers) and precision within each readability class (when using all features), defined as the number of easy or difficult sentences correctly identified as such (in their respective columns).

Accuracy was computed for all training models tested using an increasing number of features (2, 10, 30, 50 and all features) as resulting from the GRAFTING–based ranking and detailed in Table 1. Note that the first two features correspond in all cases to the traditional readability features of sentence length and word length. The classification model trained on the small gold standard dataset turned out to almost always outperform all other models: it achieved the best accuracy (83.7%) using a relatively small number of features (66), and also for a fixed number of features (i.e. 2, 30 and 50). Only when using the top–10 features, the uncorrected balanced large dataset slightly outperformed the gold standard dataset. The accuracy when using the unbalanced dataset for training was always significantly ($p < 0.05$) worse (using McNemar's test) than the accuracy based on the other training data. The only other significant difference existed between the balanced small and large dataset for 10 features. All other differences are non–significant.

It is also interesting to note that in the results reported in column *2 ft* of Table 3 a significant difference is observed when comparing the accuracy achieved using the unbalanced large data set with that achieved with the gold standard data: i.e. about $15.5$ percentage points of difference for the *2 ft* model against $3 - 6\%$ using higher numbers of features. This result originates from the fact that the unbalanced corpus contains to a larger extent sentences which are short and complex at the same time whose correct readability assessment requires linguistically–grounded features (see below).

The last two columns of Table 3 report precision results for easy– vs. difficult–to–read sentences for each of the four training datasets (all features). It is clear that for the class of difficult–to–read sentences the highest precision (88.5%) is obtained when using the whole *2Par* and *Rep* corpora for training (i.e. unbalanced large), whereas for the class of easy–to–read sentences the best precision results (84.8%) are obtained with the system trained on the gold standard dataset. Interestingly, the worst precision results (69.2%) are reported for the class of easy–to–read sentences with the unbalanced large training data set.

These results suggest that the advantages of using the gold standard data over the uncorrected training data sets are limited. From this it follows that treating the whole *Rep* corpus as a collection of difficult–to–read sentences is not completely unjustified: this is in line with the satisfactory results reported by Dell'Orletta et al. (2011) where *Rep* was used for training a sentence read-

ability classifier without any manual filtering of sentences. Nevertheless, the results of this experiment demonstrate that readability assessment accuracy and in particular the precision in identifying easy–to–read sentences can be improved by using a manually selected training dataset. Balancing the size of larger but potentially noisy (i.e. without manual revision) data sets appears to create a positive trade–off between accuracy and precision for both classes, thus representing a viable alternative to the construction of a gold standard dataset.

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

To identify the typology of features needed for sentence readability assessment and compare them to those needed for assessing document readability, we compared the results obtained by the grafting–based feature selection in the sentence classification task (using the gold standard dataset for training, see Table 3) to those obtained in the document classification task whose accuracy on the test set is reported in Table 4 for increasing numbers of features selected via GRAFTING.

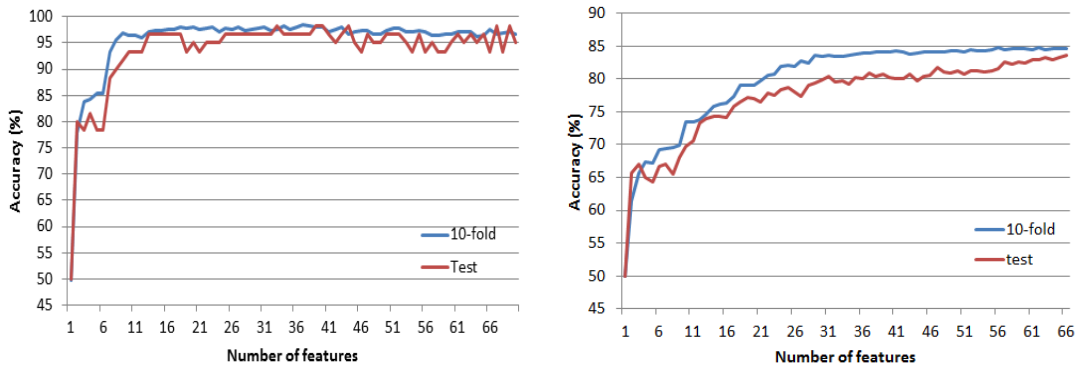
| Train. data | 2 ft | 10 ft | 30 ft | 50 ft | 70 ft (all) |
|---|---|---|---|---|---|
| *Rep - 2Par* | 80 | 93.3 | **96.6** | 96.6 | 95 |

Table 4: Accuracy of document classification for a varying number of features

By comparing the document classification results with respect to those obtained for sentences, it can be noticed that the best accuracy is achieved using a set of 30 features: in contrast to sentence classification where adding features keeps increasing the performance, more features do not appear to help for document classification. Sentence readability classification thus seems to be a more complex task, requiring a higher amount of features. This trend emerges more clearly in Figures 1(a) and 1(b), where the classification results on the training set (using 10–fold cross–validation) and the held–out test set are visualized for increasing amounts of features selected via GRAFTING. As Figure 1(a) shows, the document classification task requires about 14 features after which the performance appears to stabilize (97.4% accuracy for the ten–fold cross-validation and 96.7% for the held–out test set). In contrast, Figure 1(b) shows that sentence classification requires at least 30 features (83.4% accuracy for the ten–fold cross-validation and 79.9% for the test set).

Noticeable differences can also be observed in the typology of features playing a prominent role in the two tasks. For each feature taken into account, Table 1 reports its ranking as resulting from sentence– and document–based classification experiments (columns "Sent. class." and "Doc. class." respectively). Note that in interpreting the rank associated with each feature it should be considered that in sentence– and document–classification the number of required features is significantly different, i.e. 30 and 14 respectively: this is to say that approximately the same rank associated to the same feature does not entail a comparable role across the two classification tasks.

As already pointed out, for both sentences and documents raw text features (i.e. *Sentence length* and *Word length*) turned out to be the top features, leading however to significantly different results: i.e. 80% accuracy for documents vs. 65% for sentences. Among the remaining features, grafting results show that syntactic features do play a central role in both sentence– and document–based readability assessment: many of these are highly ranked, with some differences. Syntactic features playing a similar role in both readability classification tasks include: *Verbal root* [73], *Parse tree depth* [71], *'Chains' of embedded subordinate clauses* [83] and *Max. length of dependency links* [87], covering important aspects of syntactic complexity such as depth of the syntactic dependency (sub–)tree and length of dependency links. Features that are mainly useful for sentence readability turned out to be *Arity of verbal predicates* [74], *Pre–verbal subject* [75], *Post–verbal object* [78] and *Embedded complement 'chains'* [72], which can all be seen as representing local features referring to sentence parts. The feature *Subordinate clauses in pre–verbal position* [81], focusing on the global distribution of pre–verbal subordinate clauses within the document, is relevant for document classification only. It is interesting to note that features capturing different facets of the same phenomenon can play quite a different role for assessing the readability of sentences vs. documents: this is the case of dependency length, measured in terms of the words occurring between the syntactic head and the dependent, where feature [86] refers to the average length of all dependency links and [87] to the average length of




(a) Document classification

(b) Sentence classification

Figure 1: Document vs Sentence classification results

maximum dependency links from each sentence. Whereas [86] plays a similar role for sentences and documents (in both cases it is a middle rank feature), [87] is a global feature playing a more prominent role in document classification.

At the morpho–syntactic level, the feature ranking is more comparable. However, it is interesting to note that very few morpho–syntactic features were selected by the feature selection process: this is particularly true for document classification. This can follow from the fact that these features can be considered as proxies of the syntactic structure which in these experiments was represented through specific features: in this situation, the grafting process preferred syntactic features over morpho–syntactic ones, in spite of the lower accuracy of the dependency parser with respect to the part–of–speech tagger. Interestingly, this result is in contrast with what reported by Falkenjack and Jönsson (2014) for what concerns document readability assessment, who claim that an optimal subset of text features for readability based document classification does not need features induced via parsing. Among the morpho–syntactic features, it appears that verbal features play an important role: this can follow both by the language dealt with which is a a morphologically rich language, and by the fact that these features do not have a counterpart at the syntactic level.

Lexical features show a much more mixed result. Type–Token Ratio (TTR) 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' of the *Basic Italian Vocabulary*).

## 5 Discussion

In this study we have focused on three research questions. First, we asked which type of training 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 observed improvements were quite modest. However, we did not consider 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 (with different levels of performance), most of the features had a different rank across the two tasks, with local features being more predictive for sentence classification and global ones for documents. This suggests that the sentence readability task is more complex than assessing document readability, given that there is much less information available for a sentence than for a document.

## Acknowledgments

The research reported in this paper was carried out in the framework of the Short Term Mobility program of international exchanges funded by CNR (Italy). We thank Daniël de Kok for his help in applying TINYEST to our data and Giulia Benotto for her help in manual revision of training data.

## References

Sandra M. Aluísio, Lucia Specia, Thiago A.S. Pardo, Erick G. Maziero, and Renata P.M. Fortes. 2008. Towards brazilian portuguese automatic text simplification systems. In *Proceedings of the Eighth ACM Symposium on Document Engineering*, pages 240–248.

Sandra Aluísio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9.

Øistein E. Andersen, Helen Yannakoudakis, Fiona Barker, and Tim Parish. 2013. Developing and testing a self-assessment and tutoring system. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 32–41.

Giuseppe Attardi. 2006. Experiments with a multilanguage non-projective dependency parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X '06), New York City, New York*, pages 166–170.

Gianni Barlacchi and Sara Tonelli. 2013. Ernesta: A sentence simplification tool for children's stories in italian. In *Proceedings of the 14th Conferences on Computational Linguistics and Natural Language Processing (CICLing 2013)*, pages 476–487.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. volume 34.

Stefan Bott and Horacio Saggion. 2011. An unsupervised alignment algorithm for text simplification corpus construction. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 20–26.

Daniël de Kok. 2011. Discriminative features in reversible stochastic attribute-value grammars. In *Proceedings of the EMNLP Workshop on Language Generation and Evaluation*, pages 54–63. Association for Computational Linguistics.

Daniël de Kok. 2013. *Reversible Stochastic Attribute-Value Grammars*. Ph.D. thesis, Rijksuniversiteit Groningen.

Tullio De Mauro. 2000. *Il dizionario della lingua italiana*. Paravia, Torino.

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 Workshop on Speech and Language Processing for Assistive Technologies (SLPAT 2011)*, pages 73–83.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2014. Assessing document and sentence readability in less resourced languages and across textual genres. In *International Journal of Applied Linguistics (ITL). Special Issue on Readability and Text Simplification*. To appear.

Felice Dell'Orletta. 2009. Ensemble system for part-of-speech tagging. In *Proceedings of Evalita'09, Evaluation of NLP and Speech Tools for Italian, Reggio Emilia, December*.

Biljana Drndarević, Sanja Štajner, Stefan Bott, Susana Bautista, and Horacio Saggion. 2013. Automatic text simplification in spanish: A comparative evaluation of complementing modules. In *Computational Linguistics and Intelligent Text Processing*, pages 488–500. Springer Berlin Heidelberg.

Johan Falkenjack and Arne Jönsson. 2014. Classifying easy-to-read texts without parsing. In *Proceedings of the Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, Gothenburg, Sweden. Association for Computational Linguistics.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 276–284.

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, Jeju Island, Korea*, pages 466–477.

Michael J. Heilman, Kevyn Collins, and Jamie Callan. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of the Human Language Technology Conference*, pages 460–467.

Kentaro Inui, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida, and Tomoya Iwakura. 2003. Text simplification for reading assistance: A project note. In *Proceedings of the Second International Workshop on Paraphrasing*, pages 9–16.

Rohit J. Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J. Mooney, Salim Roukos, and Chris Welty. 2010. Learning to

predict readability using diverse linguistic features. In *oceedings of the 23rd International Conference on Computational Linguistics*, pages 546–554.

J. Peter Kincaid, Lieutenant Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas for navy enlisted personnel. In *Research Branch Report, Millington, TN: Chief of Naval Training*, pages 8–75.

Annie Louis and Ani Nenkova. 2013. A corpus of science journalism for analysing writing quality. volume 4.

Simon Perkins, Kevin Lacker, and James Theiler. 2003. Grafting: Fast, incremental feature selection by gradient descent in function space. *The Journal of Machine Learning Research*, 3:1333–1356.

Maria Emanuela Piemontese. 1996. *Capire e farsi capire. Teorie e tecniche della scrittura controllata*. Tecnodid, Napoli.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195.

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 (ACL 05)*, pages 523–530.

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

Johan Sjöholm. 2012. *Probability as readability: A new machine learning approach to readability assessment for written Swedish*. LiU Electronic Press, Master thesis.

Adam Skory and Maxine Eskenazi. 2010. Predicting cloze task quality for vocabulary training. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 49–56.

Sara Tonelli, Ke Tran Manh, and Emanuele Pianta. 2012. Making readability indices readable. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 40–48.

Sowmya Vajjala and Detmar Meurers. 2014. On assessing the reading level of individual sentences 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.

Sanja Štajner and Horacio Saggion. 2013. Readability indices for automatic evaluation of text simplification systems: A feasibility study for spanish. In *Proceedings of the International Joint Conference on Natural Language Processing*.