READ–IT:
Assessing Readability of Italian Texts with a View to Text Simplification

Felice Dell'Orletta, Simonetta Montemagni and Giulia Venturi
Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
via G. Moruzzi, 1 – Pisa (Italy)
{felice.dellorletta,simonetta.montemagni,giulia.venturi}@ilc.cnr.it

Abstract

In this paper, we propose a new approach to readability assessment with a specific view to the task of text simplification: the intended audience includes people with low literacy skills and/or with mild cognitive impairment. READ–IT represents the first advanced readability assessment tool for what concerns Italian, which combines traditional raw text features with lexical, morpho-syntactic and syntactic information. In READ–IT readability assessment is carried out with respect to both documents and sentences where the latter represents an important novelty of the proposed approach creating the prerequisites for aligning the readability assessment step with the text simplification process. READ–IT shows a high accuracy in the document classification task and promising results in the sentence classification scenario.

1 Introduction

Recently, there has been increasing interest in the exploitation of results from Natural Language Processing (NLP) for the development of assistive technologies. Here, we address this topic by reporting the first but promising results in the development of a software architecture for the Italian language aimed at assisting people with low literacy skills (both native and foreign speakers) or who have language disabilities in reading texts.

Within an information society, where everyone should be able to access all available information, improving access to written language is becoming more and more a central issue. This is the case, for instance, of administrative and governmental information which should be accessible to all members of the society, including people who have reading difficulties for different reasons: because of a low education level or because of the fact that the language in question is not their mother tongue, or because of language disabilities. Health related information represents another crucial domain which should be accessible to a large and heterogenous target group. Understandability in general and readability in particular is also an important issue for accessing information over the web as stated in the Web Content Accessibility Guidelines (WCAG) proposed by the Web Accessibility Initiative of the W3C.

In this paper, we describe the approach we developed for automatically assessing the readability of newspaper texts with a view to the specific task of text simplification. The paper is organized as follows: Section 2 describes the background literature on the topic; Section 3 introduces the main features of our approach to readability assessment, with Section 4 illustrating its implementation in the READ-IT prototype; Sections 5 and 6 describe the experimental setting and discuss achieved results.

2 Background

Readability assessment has been a central research topic for the past 80 years which is still attracting considerable interest nowadays. Over the last ten years, within the NLP community the automatic assessment of readability has received increasing attention: if on the one hand the availability of sophisticated NLP technologies makes it possible to monitor a wide variety of factors affecting the readability

of a text, on the other hand there is a wide range of both human- and machine-oriented applications which can benefit from it.

Traditional readability formulas focus on a limited set of superficial text features which are taken as rough approximations of the linguistic factors at play in readability assessment. For example, the Flesch-Kincaid measure (the most common reading difficulty measure still in use, Kincaid (1975)) is a linear function of the average number of syllables per word and of the average number of words per sentence, where the former and latter are used as simple proxies for lexical and syntactic complexity respectively. For Italian, there are two readability formulas: an adaptation of the Flesh-Kincaid for English to Italian known as the Flesch-Vacca formula (Franchina and Vacca, 1986); the GulpEase index (Lucisano and Piemontese, 1988), assessing readability on the basis of the average number of characters per word and the average number of words per sentence.

A widely acknowledged fact is that all traditional readability metrics are quick and easy to calculate but have drawbacks. For example, the use of sentence length as a measure of syntactic complexity assumes that a longer sentence is more grammatically complex than a shorter one, which is often but not always the case. Word syllable count is used starting from the assumption that more frequent words are more likely to have fewer syllables than less frequent ones (an association that is related to Zipf's Law, Zipf (1935)); yet, similarly to the previous case, word length does not necessarily reflects its difficulty. The unreliability of these metrics has been experimentally demonstrated by several recent studies in the field: to mention only a few Si and Callan (2001), Petersen and Ostendorf (2006), Feng (2009).

On the front of the assessment of the lexical difficulty of a given text, a first step forward is represented by vocabulary-based formulas such as the Dale-Chall formula (Chall and Dale, 1995), using a combination of average sentence length and word frequency counts. In particular, for what concerns the latter it reconstructs the percentage of words not on a list of 3000 "easy" words by matching its own list to the words in the material being evaluated, to determine the appropriate reading level. If vocabulary-based measures represent an improvement in assessing the readability of texts which was possible due to the availability of frequency dictionaries and reference corpora, they are still unsatisfactory for what concerns sentence structure.

Over the last ten years, work on readability deployed sophisticated NLP techniques, such as syntactic parsing and statistical language modeling, to capture more complex linguistic features and used statistical machine learning to build readability assessment tools. A variety of different NLP-based approaches to the automatic readability assessment has been proposed so far, differing with respect to: a) the typology of features taken into account (e.g. lexical, syntactic, semantic, discourse), and, for each type, at the level of the inventory of used individual features; b) the intended audience of the texts under evaluation, which strongly influences the readability assessment, and last but not least c) the application within which readability assessment is carried out.

Interesting alternatives to static vocabulary-based measures have been put forward by Si and Callan (2001) who used unigram language models combined with sentence length to capture content information from scientific web pages, or by Collins-Thompson and Callan (2004) who adopted a similar language modeling approach (Smoothed Unigram model) to predict reading difficulty of short passages and web documents. These approaches can be seen as a generalization of the vocabulary-based approach, aimed at capturing finer-grained and more flexible information about vocabulary usage. If unigram language models help capturing important content information and variation of word usage, they do not cover other types of features which are reported to play a significant role in the assessment of readability. More recently, the role of syntactic features started being investigated (Schwarm and Ostendorf, 2005; Heilman et al., 2007; Petersen and Ostendorf, 2009): in these studies syntactic structure is tracked through a combination of features from n-gram (trigram, bigram and unigram) language models and parse trees (parse tree height, number of noun phrases, verb phrases and subordinated clauses or SBARs) with more traditional features.

Yet, besides lexical and syntactic complexity features there are other important factors, such as the structure of the text, the definition of discourse topic, discourse cohesion and coherence and so on, play-

ing a central role in determining the reading difficulty of a text. More recent approaches esplored the role of these features in readability assessment: this is the case, for instance, of Barzilay and Lapata (2008) or Feng (2010). The last few years have been characterised by approaches based on the combination of features ranging over different linguistic levels, namely lexical, syntactic and discourse (see e.g. Pitler and Nenkova (2008), Kate (2010)).

Another important factor determining the typology of features to be considered for assessing readability has to do with the intended audience of readers: it is commonly agreed that reading ease does not follow from intrinsic text properties alone, but it is also affected by the expected audience. Among the studies addressing readability with respect to specific audiences, it is worth mentioning here: Schwarm and Ostendorf (2005) and Heilman et al. (2007) dealing with language learners, or Feng (2009) focussing on people with mild intellectual disabilities. Interestingly,Heilman et al. (2007) differentiate the typology of used features when addressing first (L1) or second (L2) language learners: they argue that grammatical features are more relevant for L2 than for L1 learners. Feng (2009) propose a set of cognitively motivated features operating at the discourse level specifically addressing the cognitive characteristics of the expected users. When readability is targeted towards adult competent language users a more prominent role is played by discourse features (Pitler and Nenkova, 2008).

Applications which can benefit from an automatic readability assessment range from the selection of reading material tailored to varying literacy levels (e.g. for L1/L2 students or low literacy people) and the ranking of documents by reading difficulty (e.g. in returning the results of web queries) to NLP tasks such as automatic document summarization, machine translation as well as text simplification. Again, also the application making use of the readability assessment, which is in turn strictly related to the intended audience of readers, strongly influences the typology of features to be taken into account.

Advanced NLP–based readability metrics developed so far typically deal with English, with a few attempts devoted to other languages, namely French (Collins-Thompson and Callan, 2004), Portuguese (Aluisio et al., 2010) and German (Brück, 2008).

3 Our Approach

Our approach to readability assessment was developed with a specific application in mind, i.e. text simplification, and addresses a specific target audience of readers, namely people characterised by low literacy skills and/or by mild cognitive impairment. Following the most recent approaches, we treat readability assessment as a classification task: in particular, given the available corpora for the Italian language as well as the type of target audience, we resorted to a binary classification aimed at discerning easy–to–read textual objects from difficult–to–read ones. The language dealt with is Italian: to our knowledge, this is the first attempt of an advanced methodology for readability assessment for this language. Our approach focuses on lexical and syntactic features, whose selection was influenced by the application, the intended audience and the language dealt with (both for its intrinsic linguistic features and for the fact of being a less resourced language). Following Roark (2007), in the features selection process we preferred easy-to-identify features which could be reliably identified within the output of NLP tools. Last but not least, as already done by Aluisio et al. (2010) the set of selected syntactic features also includes simplification oriented ones, with the final aim of aligning the readability assessment step with the text simplification process.

Another qualifying feature of our approach to readability assessment consists in the fact that we are dealing with two types of textual objects: documents and sentences. The latter represents an important novelty of our work since so far most research focused on readability classification at the document level (Skory and Eskenazi, 2010). When the target application is text simplification, we strongly believe that also assessing readability at the sentence level could be very useful. We know that methods developed so far perform well to characterize the level of an entire document, but they are unreliable for short texts and thus also for single sentences. Sentence-based readability assessment thus represents a further challenge we decided to tackle: in fact, 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 difficult-to-read sen-

tences. Since there are no training data at the sentence level, it becomes difficult – if not impossible – to evaluate the effectiveness of our approach, 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. In order to overcome this problem in the readability assessment of individual sentences, we introduced a notion of distance with respect to easy-to-read sentences. In this way, the prerequisites are created for the integration of the two processes of readability assessment and text simplification. Before, text readability was assessed with respect to the entire document and text simplification was carried out at the sentence level: due to the decoupling of the two processes, the impact of simplification operations on the overall readability level of the text was not always immediately clear. With sentence-based readability assessment, this should be no longer a problem.

4 READ–IT

Our approach to readability assessment has been implemented in a software prototype, henceforth referred to as READ–IT. READ–IT operates on syntactically (i.e. dependency) parsed texts and it assigns to each considered reading object - either a document or a sentence - a score quantifying its readability. READ–IT is a classifier based on Support Vector Machines using LIBSVM (Chang and Lin, 2001) that, given a set of features and a training corpus, creates a statistical model using the feature statistics extracted from the training corpus. Such a model is used in the assessment of readability of unseen documents and sentences.

The set of features used to build the statistical model can be parameterized through a configuration file: as we will see, the set of relevant features used for readability assessment at the document level differs from the those used at the sentence level. This also creates the prerequisites for specialising the readability assessment measure with respect to more specific target audiences: as pointed out in Heilman et al. (2007) different types of features come into play e.g. when addressing L1 or L2 language learners. Here follows the complete list of features used in the reported experiments.

4.1 Features

The features used for predicting readability are organised into four main categories: namely, raw text features, lexical features as well as morpho-syntactic and syntactic features. This proposed four–fold partition closely follows the different levels of linguistic analysis automatically carried out on the text being evaluated, i.e. tokenization, lemmatization, PoS tagging and dependency parsing. Such a partition was meant to identify those easy to extract features with high discriminative power in order to reduce the linguistic pre-processing of texts guaranteeing at the same time a reliable readability assessment.

Raw Text Features

They 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: 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); *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, it can be seen as the complement of the type out-of-vocabulary rate features used by Petersen and Ostendorf (2009).

Type/Token Ratio: this feature refers to the ratio between the number of lexical types and the number of tokens. This feature, which can be considered as an indicator of expressive language delay or

disorder as shown in Wright (2003) for adults and in Retherford (2003) for children, has already been used for readability assessment purposes by Aluisio et al. (2010). Due to its sensitivity to sample size, this feature has been computed for text samples of equivalent length.

Morpho–syntactic Features

Language Model probability of Part-Of-Speech unigrams: 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 types (POS) and their individual probabilities. This feature has already been shown to be a reliable indicator for automatic readability assessment (see, for example, Pitler and Nenkova (2008) and Aluisio et al. (2010)).

Lexical density: 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 already been used for readability assessment by Aluisio et al. (2010) and Feng (2010).

Verbal mood: this feature refers to the distribution of verbs according to their mood. It is a novel and language–specific feature exploiting the predictive power of the Italian rich verbal morphology.

Syntactic Features

Unconditional probability of dependency types: 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 (2010).

Parse tree depth features: parse tree depth can be indicative of increased sentence complexity as stated by, to mention only a few, 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; b) the *average depth of embedded complement 'chains'* governed by a nominal head and including either prepositional complements or nominal and adjectival modifiers; c) the *probability distribution of embedded complement 'chains' by depth*. The first feature has already been used in syntax-based readability assessment studies (Schwarm and Ostendorf, 2005; Heilman et al., 2007; Nenkova, 2010); the latter two are reminiscent of the 'head noun modifiers' feature used by Nenkova (2010).

Verbal predicates features: 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. The *arity of verbal predicates* is calculated as the number of instantiated dependency links sharing the same verbal head (covering both arguments and modifiers). Although there is no obvious relation between the number of verb dependents and sentence complexity, we believe that both a low and a high number of dependents can make sentence readability quite complex, although for different reasons (elliptical constructions in the former case, a high number of modifiers in the latter). Within this feature set we also considered the *distribution of verbal predicates by arity*. To our knowledge, this set of features has never been used so far for readability assessment purposes.

Subordination features: subordination is widely acknowledged to be an index of structural complexity in language. As in Aluisio et al. (2010), this set of features has been introduced here with a specific view to the text simplification task. A first feature was meant to measure the *distribution of subordinate vs main clauses*. For subordinates, we also considered 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. Two further features were introduced to capture the depth of embedded subordinate clauses since it is a widely acknowledged fact that highly complex sentences contain deeply embedded subordinate clauses: in particular, a) the *average depth of 'chains' of embedded subordinate clauses* and b) the *probability distribution of embedded subordinate clauses 'chains' by depth*.

Length of dependency links feature: 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 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 (2010) and Feng (2010).

5 The Corpora

One challenge in this work was finding an appropriate corpus. Although a possibly large collection of texts labelled with their target grade level (such as the Weekly Reader for English) would be ideal, we are not aware of any such collection that exists for Italian in electronic form. Instead, to test our approach to automatically identify the readability of a given text, we used two different corpora: a newspaper corpus, *La Repubblica* (henceforth, "Rep"), and an easy–to–read newspaper, *Due Parole* (henceforth, "2Par") which was specifically written for an audience of adults with a rudimentary literacy level or with mild intellectual disabilities. The articles in 2Par were written by Italian linguists expert in text simplification using a controlled language both at the lexicon and sentence structure levels (Piemontese, 1996) .

There are different motivations underlying the selection of these two corpora for our study. On the practical side, to our knowledge 2Par is the only available corpus of simplified texts addressing a wide audience characterised by a low literacy level. So, the use of 2Par represented the only possible option on the front of simplified texts. For the selection of the second corpus we opted for texts belonging to the same class, i.e. newspapers: this was aimed at avoiding interferences due to textual genre variation in the measure of text readability. This is confirmed by the fact that the two corpora show a similar behaviour with respect to a number of different parameters, which according to the literature on register variation (Biber, 2009) are indicative of textual genre differences: e.g. lexical density, the noun/verb ratio, the percentage of verbal roots, etc. On the other hand, the two corpora differ significantly with respect to the distribution of features typically correlated with text complexity, e.g. the composition of the used vocabulary (e.g. the percentage of words belonging to the *Basic Italian Vocabulary* in Rep is 4.14% and in 2Par is 48.04%) or, from the syntactic point of view, the average parse tree height (which in Rep is 5.71 and in 2Par 3.67), the average number of verb phrases per sentence (which in Rep is 2.40 and in 2Par 1.25), the depth of nested structures (e.g. the average depth of embedded complement 'chains' in Rep is 1.44 and in 2Par is 1.30), the proportion of main vs subordinate clauses (in Rep main and subordinate clauses represent respectively 65.11% and 34.88% of the cases; in 2Par there is 79.66% of main clauses and 20.33% of subordinate clauses).

The Rep/2Par pair of corpora is somehow reminiscent of corpora used in other readability studies, such as Encyclopedia Britannica and Britannica Elementary, but with a main difference: whereas the English corpora consist of paired original/simplified texts, which we might define as "parallel monolingual corpora", the selected Italian corpora rather present themselves as "comparable monolingual corpora", without any pairing of the full–simplified versions of the same article. Comparability is guaranteed here by the inclusion of texts belonging to the same textual genre: we expect such comparable corpora to be usefully exploited for readability assessment because of the emphasis on style over topic.

Although these corpora do not provide an explicit grade–level ranking for each article, broad categories are distinguished, namely easy–to–read vs difficult–to–read texts. The two paired complex/simplified corpora were used to train and test different language models described in Section 6. As already pointed out, such a distinction is reliable in a document classification scenario, while at the sentence classification level it poses the remarkable issue of discerning easy–to–read sentences within difficult–to–read documents (i.e. Rep).

6 Experiments and Results

READ–IT was tested on the 2Par and Rep corpora 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 Machine as learning algorithm. Three different sets of experiments were devised to test the performance of READ-IT in the following subtasks: i) document readability classification, ii) sentence readability classification and iii) detection of easy–to–read sentences within difficult–

to–read texts.

For what concerns the document classification subtask, we used a corpus made up of 638 documents of which 319 were extracted from 2Par (taken as representative of the class easy–to–read texts) and 319 from Rep (representing the class of difficult–to–read texts). We have followed a *5*–fold cross–validation process: the corpus was randomly split into 5 training and test sets. The test sets consisted of 20% of the individual documents belonging to the two considered readability levels, with each document being included in one test set only. With regard to the sentence classification subtask, we used a training set of about 3,000 sentences extracted from 2Par and of about 3,000 sentences from Rep and a test corpus of 1,000 sentences of which 500 were extracted from 2Par (hereafter, *2Par test set*) and 500 from Rep (hereafter, *Rep test set*). In the third experiment, readability assessment was carried out by READ–IT with respect to a much bigger corpus of 2,5 milion of words extracted from the newspaper *La Repubblica* (hereafter, *Rep 2.5*), for a total of 123,171 sentences, with the final aim of detecting easy–to–read sentences.

All the experiments were carried out using four different readability models, described as follows:

1. **Base Model**, using *raw text* features only;
2. **Lexical Model**, using a combination of *raw text* and *lexical* features;
3. **MorphoS Model**: using *raw text*, *lexical* and *morpho–syntactic* features;
4. **Syntax Model**: combining all feature types, namely *raw text*, *lexical*, *morpho–syntactic* and *syntactic* features.

Note that in the Lexical and Syntax Models, different sets of features were selected for the subtasks of document and sentence classification. In particular, for sentence–based readability assesment we did not take into account the Type/Token Ratio feature, all features concerning the distribution of 'chains' of embedded complements and subordinate clauses and the distribution of verbal predicates by arity.

Since, to our knowledge, a machine learning readability classifier does not exist for the Italian language we consider the *Base Model* as our baseline: this can be seen as an approximation of the GulpEase index, which is based on the same raw text features (i.e. sentence and word length).

6.1 Evaluation Methodology

Different evaluation methods have been defined in order to assess achieved results in the three aforementioned experiment sets. The performance of both document and sentence classification experiments have been evaluated in terms of i) overall Accuracy of the system and ii) Precision and Recall.

In particular, Accuracy is a global score referring to the percentage of documents or sentences correctly classified, either as easy–to–read or difficult–to–read objects. Precision and Recall have been computed with respect to two the target reading levels: in particular, Precision is the ratio of the number of correctly classified documents or sentences over the total number of documents and sentences classified by READ–IT as belonging to the easy–to–read (i.e. 2Par) or difficult–to–read (i.e. Rep) classes; Recall has been computed as the ratio of the number of correctly classified documents or sentences over the total number of documents or sentences belonging to each reading level in the test sets. For each set of experiments, evaluation was carried out with respect to the four models of the classifier.

Following from the assumption that 2Par contains only easy–to–read sentences while Rep does not necessarily contain only difficult–to–read ones, we consider READ–IT errors in the classification of 2Par sentences as erroneously classified sentences. On the other hand, classification errors within the set of Rep sentences deserve an in–depth error analysis, since we need to discern real errors from misclassifications due to the fact that we are in front of easy–to–read sentences occurring in a difficult–to–read context. In order discern errors from 'correct' misclassifications, we introduced a new evaluation methodology, based on the notion of *Euclidean distance* between feature vectors. Each feature vector is a n–dimensional vector of linguistic features (see Section 4.1) that represents a set of sentences. Two vectors with 0 distance represent the same set of sentences, i.e. those sentences sharing the same values for the monitored linguistic features. Conversely, the bigger the distance between two vectors is, the more distant are the two represented sets of

sentences with respect to the monitored features.

The same notion of distance has also been used to test which model was more effective in predicting the readability of n–word long sentences.

6.2 Results

In Table 1, the Accuracy, Precision and Recall scores achieved with the different READ–IT models in the document classification subtask are reported. It can be noticed that the *Base Model* shows the lowest performance, while the *MorphoS Model* outperforms all other models. Interestingly, the *Lexical Model* shows a high accuracy for what concerns the document classification subtask (95.45%), by significantly improving the accuracy score of the *Base Model* (about +19%). This result demonstrates that for assessing the readability of documents a combination of raw and lexical features provides reliable results which can be further improved (about +3%) by also taking into account morpho-syntax.

		2Par		*Rep*	
Model	Accuracy	Prec	Rec	Prec	Rec
Base	76.65	74.71	80.56	78.91	72.73
Lexical	95.45	95.60	95.30	95.31	95.61
MorphoS	**98.12**	**98.12**	**98.12**	**98.12**	**98.12**
Syntax	97.02	97.17	96.87	96.88	97.18

Table 1: Document classification results

Consider now the sentence classification subtask. Table 2 shows that in this case the most reliable results are achieved with the *Syntax Model*. It is interesting to note that the morpho–syntactic and syntactic features allow a much higher increment in terms of Accuracy, Precision and Recall scores than in the document classification scenario: i.e. the difference between the performance of the *Lexical Model* and the best one in the document classification experiment (i.e. the *MorphoS Model*) is equal to 2.6%, while in the sentence classification case (i.e. *Syntax Model*) is much higher, namely 17% .

In Table 3, we detail the performance of the best READ–IT model (i.e. the *Syntax Model*) on the *Rep test set*. In order to evaluate those sentences which were erroneously classified as belonging to 2Par, we calculated the distance between 2Par and i) these sentences (140 sentences referred to as *wrong* in the Table), ii) the correctly classified sentences

		2Par		*Rep*	
Model	Accuracy	Prec	Rec	Prec	Rec
Base	59.6	55.6	95.0	82.9	24.2
Lexical	61.6	57.3	91.0	78.1	32.2
MorphoS	76.1	72.8	83.4	80.6	68.8
Syntax	**78.2**	**75.1**	**84.4**	**82.2**	**72.0**

Table 2: Sentence classification results

(360 sentences, referred to as *correct* in the Table), iii) the whole *Rep test set*. As we can see, the distance between the *wrong* sentences and 2Par is much lower than the distance holding between 2Par and the correcly classified sentences (*correct*). This entails that the sentences which were erroneously classified as easy–to–read sentences (i.e. belonging to 2Par) are in fact more readable than the correctly classified ones (as belonging to Rep).It is obvious that the *Rep test set*, which contains both *correct* and *wrong* sentences, has an intermediate distance value with respect to 2Par.

	Distance
Correct	52.072
Rep test set	45.361
Wrong	37.843

Table 3: Distances between 2Par and Rep on the basis of the *Syntax Model*

In Table 4, the percentage of *Rep 2.5* sentences classified as difficult–to–read is reported. The results show that the *Syntax Model* classifies the higher number of sentences as difficult–to–read, but from these results we cannot say whether this model is the best one or not since *Rep 2.5* sentences are not annotated with readability information. Therefore, in order to compare the performance of the four READ–IT models and to identify which is the best one, we computed the distance between the sentences classified as easy–to–read and 2Par, which is reported, for each model, in Table 5. It can be noticed that the *Syntax Model* appears to be the best one since it shows the lowest distance with respect to 2Par; on the other hand, the whole *Rep 2.5* corpus shows a higher distance since it contains both difficult– and easy–to–read sentences. Obviously, the sentences classified as difficult–to–read by the *Syntax Model* (*Diff Syntax* in the Table) show the broader distance.

	Accuracy
Base	0.234
Lexical	0.387
MorphoS	0.705
Syntax	**0.755**

Table 4: Accuracy in sentence classification of *Rep 2.5*.

	Distance
Diff Syntax	66.526
Rep 2.5	64.040
Base	61.135
Lexical	60.529
MorphoS	55.535
Syntax	**51.408**

Table 5: Distance between 2Par and i) difficult–to–read sentences according to the *Syntax Model*, ii) *Rep 2.5*, iii) easy–to–read sentences by the four models.

In order to gain an in–depth insight into the different behaviour of the four READ–IT models, we evaluated their performances for sentences of a fixed length. We considered sentences whose length ranges between 8 and 30. For every set of sentences of the same legth, we compared the easy–to-read sentences of *Rep 2.5* classified by the four models with respect to 2Par. In Figure 1, each point represents the distance between a set of sentences of the same length and the same n–word long set of sentences in the 2Par corpus. As it can be seen, the bottom line which represents the sentences classified as easy–to–read by the *Syntax Model* is the closest to the 2Par sentences of the same length. On the contrary, the line representing the sentences classified by the *Base Model* is the most distant amongst the four READ–IT models. Interestingly, it overlaps with the line representing the *Rep 2.5* sentences: this suggests that a classification model based only on raw text features (i.e. sentence and word length) is not able to identify easy–to–read sentences if we consider sets of sentences of a fixed length. Obviously, the line representing the sentences classified as difficult–to–read by the *Syntax Model* shows the broadest distance. This experiment has shown that linguistically motivated features (and in particular syntactic ones) have a fundamental role in the sentence readability assessment subtask.

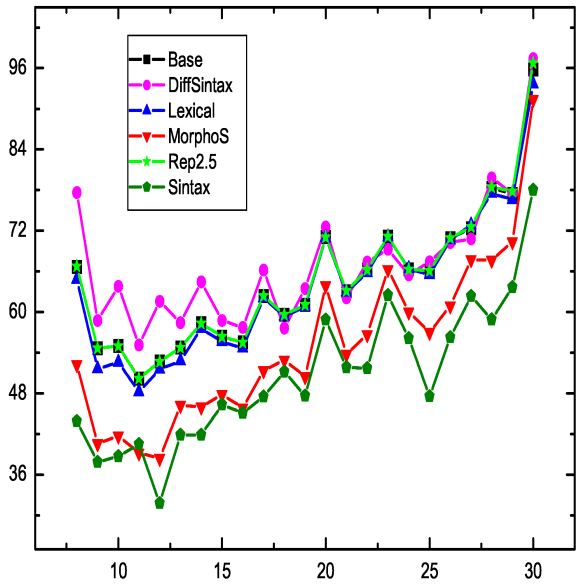

Figure 1: Distance between 2Par and i) difficult–to–read sentences according to the *Syntax Model*, ii) *Rep 2.5*, iii) easy–to–read sentences by the four models for sets of sentences of fixed length

7 Conclusion

In this paper, we proposed a new approach to readability assessment with a specific view to the task of text simplification: the intended audience includes people with low literacy skills and/or with mild cognitive impairment. The main contributions of this work can be summarised as follows: i) READ–IT represents the first advanced readability assessment tool for what concerns Italian; ii) it combines traditional raw text features with lexical, morpho-syntactic and syntactic information; iii) readability assessment is carried out with respect to both documents and sentences. Sentence–based readability assessment is an important novelty of our approach which creates the prerequisites for aligning readability assessment with text simplification. READ–IT shows a high accuracy in the document classification task and promising results in the sentence classification scenario. The two different tasks appear to enforce different requirements at the level of the underlying linguistic features. To overcome the lack of an Italian reference resource annotated with readability information at the sentence level we introduced the notion of distance to assess READ–IT performance.

Acknowledgements

The research reported in the paper has been partly supported by the national project "Migrations" of the National Council of Research (CNR) in the framework of the line of research *Advanced technologies and linguistic and cultural integration in the school*. In particular the authors would like to thank Eva Maria Vecchi who contributed to the prerequisites of the proposed readability assessment methodology reported in the paper.

References

Sandra Aluisio, 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, pp. 1–9.

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, pp. 166–170.

Regina Barzilay and Mirella Lapata. 2008. *Modeling local coherence: An entity-based approach*. In Computational Linguistics, 34(1), pp. 1–34.

Douglas Biber and Susan Conrad. 2009. *Register, genre, and style*. Cambridge, Cambridge University Press.

Tim vor der Brück, Sven Hartrumpf, Hermann Helbig. 2008. *A Readability Checker with Supervised Learning using Deep Syntactic and Semantic Indicators*. In Proceedings of the 11th International Multiconference: Information Society - IS 2008 - Language Technologies, Ljubljana, Slovenia, pp. 92–97.

Jeanne S. Chall and Edgar Dale. 1995. *Readability Revisited: The New Dale–Chall Readability Formula*. Brookline Books, Cambridge, MA.

Chih-Chung Chang and Chih-Jen Lin. 2001. *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

Kevyn Collins-Thompson and Jamie Callan. 2004. *A language modeling approach to predicting reading difficulty*. In Proceedings of the HLT / NAACL, pp. 193–200.

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.

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

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), pp. 276–284.

Lijun Feng, Noémie Elhadad and Matt Huenerfauth. 2009. *Cognitively motivated features for readability assessment*. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09), pp. 229–237.

V. Franchina and Roberto Vacca. 1986. *Adaptation of Flesh readability index on a bilingual text written by the same author both in Italian and English languages*. In Linguaggi (3), pp. 47–49.

Lyn Frazier. 1985. *Syntactic complexity*. In D.R. Dowty, L. Karttunen and A.M. Zwicky (eds.), *Natural Language Parsing*, Cambridge University Press, Cambridge, UK.

Edward Gibson. 1998. *Linguistic complexity: Locality of syntactic dependencies*. In *Cognition*, 68(1), pp. 1-76.

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, pp. 460–467.

Michael J. Heilman, Kevyn Collins and Maxine Eskenazi. 2008. *An analysis of statistical models and features for reading difficulty prediction*. In Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications (EANL '08), pp. 71–79.

Rohit J. Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J. Mooney, Salim Roukos, Chris Welty. 2010. *Learning to Predict Readability using Diverse Linguistic Features*. In Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), pp. 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*. Research Branch Report, Millington, TN: Chief of Naval Training, pp. 8–75.

George Kingsley Zipf. 1988. *The Psychobiology of Language*. Houghton–Miflin, Boston.

Dekan Lin. 1996. *On the structural complexity of natural language sentences*. In Proceedings of COLING 1996, pp. 729–733.

Pietro Lucisano and Maria Emanuela Piemontese. 1988. *Gulpease. Una formula per la predizione della difficoltà dei testi in lingua italiana*. In Scuola e Città (3), pp. 57–68.

Ryan McDonald and Joakim Nivre. 2007. *Characterizing the Errors of Data-Driven Dependency Parsing Models*. In Proceedings of EMNLP-CoNLL, 2007, pp. 122-131.

Jim Miller and Regina Weinert. 1998. *Spontaneous spoken language. Syntax and discourse*. Oxford, Clarendon Press.

Ani Nenkova, Jieun Chae, Annie Louis, and Emily Pitler. 2010. *Structural Features for Predicting the Linguistic Quality of Text Applications to Machine Translation, Automatic Summarization and Human–Authored Text*. In E. Krahmer, M. Theune (eds.), *Empirical Methods in NLG*, LNAI 5790, Springer-Verlag Berlin Heidelberg, pp. 222–241.

Sarah E. Petersen and Mari Ostendorf. 2006. *A machine learning approach to reading level assessment*. University of Washington CSE Technical Report.

Sarah E. Petersen and Mari Ostendorf. 2009. *A machine learning approach to reading level assessment*. In Computer Speech and Language (23), pp. 89–106.

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

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, pp. 186–195.

Kristine Retherford. 2003. *Normal development: a database of communication and related behaviors*. Eau Claire, WI: Thinking Publications.

Brian Roark, Margaret Mitchell and Kristy Hollingshead. 2007. *Syntactic complexity measures for detecting mild cognitive impairment*. In Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, pp. 1–8.

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), pp. 523–530.

Luo Si and Jamie Callan. 2001. *A statistical model for scientific readability*. In Proceedings of the tenth international conference on Information and knowledge management, pp. 574–576.

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, pp. 49–56.

Victor H.A. Yngve. 1960. *A model and an hypothesis for language structure*. In Proceedings of the American Philosophical Society, pp. 444-466.

Heather Harris Wright, Stacy W. Silverman and Marilyn Newhoff. 2003. *Measures of lexical diversity in aphasia*. In Aphasiology, 17(5), pp. 443-452.