Assessing Document and Sentence Readability in Less Resourced Languages and across Textual Genres

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 tackle three open issues of the automatic readability assessment literature, namely: the evaluation of text readability in less resourced languages, with respect to sentences (as opposed to documents) as well as across textual genres. Different solutions to these issues have been tested by using and specialising READ–IT, 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: READ–IT shows a high accuracy in the document classification task and promising results in the sentence classification scenario. By comparing the results of two versions of READ-IT, adopting a classification- *vs* ranking-based approach, we also show that readability assessment is strongly influenced by textual genre and for this reason a genre–oriented notion of readability is needed. With classification–based approaches, reliable results can only be achieved with genre–specific models: since this is far from being a workable solution, especially for less–resourced languages, a new ranking method for readability assessment is proposed, based on the notion of distance.

Key words

Classification; Less Resourced Languages; Multi-Level Linguistic Annotation; Ranking; Readability; Textual Genres

1 Introduction

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 (as in the case of immigrants), 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* (Caldwell et al., 2008) proposed by the Web Accessibility Initiative of the W3C.[1] Last but not least, education is another area within which readability assessment plays a central role by enabling the selection of appropriate reading material for readers of varying proficiency: providing students with texts that are accessible and well matched to reader abilities has always been a challenge for educators.

The increasigly acknowledged potential of readability assessment in the wide range of contexts listed above combined with the development of efficient and sophisticated Natural Language Processing (NLP) techniques led recently to a resurgence of interest in automatic readability assessment. Over the last ten years, several studies have been carried out within the Computational Linguistics community deploying NLP techniques to capture a wide range of multi-level linguistic (e.g. lexical, syntactic, discourse) features and using statistical machine learning to build advanced readability assessment tools. These studies show that NLP-enabled feature extraction and state–of–the–art machine learning algorithms provide significantly improved results with respect to traditional readability measures.

NLP–based approaches to readability assessment proposed in the literature can be subdivided into two groups, according to whether readability assessment is carried out as a classification task (see, among others, Petersen and Ostendorf, 2009, Aluisio et al., 2010; Feng et al., 2010; Nenkova et al., 2010) or in terms of ranking (see, among others, Tanaka-Ishii et al., 2010; Ma et al., 2012; Inui and Yamamoto, 2001). Methods following a classification approach carry out this task by assigning the document under analysis to a specific readability class, while ranking–based methods assign the document a score positioning it within a readability ranking scale. From this it follows that whereas a classification–based

[1] http://www.w3.org/WAI/

system requires a predefined set of classes of readability, ranking methods do not assume any readability leveling system besides the extreme poles representing maximum and minimum readability.

Most part of the NLP-based tools developed so far treat readability assessment as a classification task: the main problem of this type of methods is represented by the lack of training data representative of fine-grained readability classes. It goes without saying that this is even more problematic for languages which are less–resourced, at least as far as readability is concerned: this is the case, for instance, of Italian for which we are not aware of the existence of any large collections of digital texts labelled with an articulated set of target grade levels. Ranking–based readability assessment methods represent a viable alternative to classification methods, since they only require training data with respect to two readability levels (i.e. easy *vs* difficult). Ranking-based approaches can also be of some help when finer-grained and/or personalised levels of readability are required, as in the context of e.g. self-directed language learning (Beinborn et al., 2012).

As pointed out by Skory and Eskenazi (2010), most research focused on readability assessment at the document level: i.e. methods developed so far perform well to characterize the level of an entire document, but they are unreliable for short texts and therefore also for single sentences. However, for specific applicative purposes assessing the readability level of individual sentences would also be desirable. Consider, for instance, the case of text simplification: in the approaches proposed so far, text readability is typically assessed with respect to the entire document and text simplification is instead carried out at the sentence level (as e.g. in Aluisio et al., 2010): due to the decoupling of the two processes, the impact of simplification operations on the overall readability level of the text may not always be immediately clear. With sentence-based readability assessment, this should be no longer a problem. Sentence-based readability assessment thus represents an open issue which in our opinion should be further explored for the new generation of NLP-based redability assessment tools to be used for driving the text simplification process, whenever required.

A further open issue emerges from the most recent literature which reports that the degree of readability is, at least to some extent, connected to the textual genre of the document under evaluation: consider, for instance, the work by Kate et al. (2010) who improves the accuracy of readability predictions by using genre–specific features, or by Štajner et al. (2012) who proved that linguistic features correlated with readability are also genre dependent. This suggests that textual genre and readability do not represent orthogonal dimensions of classification, but intertwined notions whose complex interplay needs to be further investigated in order to envisage solutions which could be successfully exploited in real applications. In principle, texts to be evaluated for readability can belong to different genres, ranging e.g. from fiction to scientific writing or reportage. The question which naturally arises is then whether and to what extent readability assessment is genre–independent, and if this is not the case whether and how general purpose readability assessment tools could reliably be used for dealing with texts belonging to different genres.

On the basis of what was said so far, evaluation of readability for less resourced languages, with respect to different types of reading objects (documents *vs* sentences) as well as across textual genres, still represents a set of open issues in the area of automatic readability assessment which needs to be further explored: the work reported in this paper is aimed at shedding light on them, by highlighting the problems of current readability assessment methods and techniques and by proposing innovative solutions to overcome detected gaps and limits. Our solutions to the issues addressed in this paper have been tested on Italian, for which very few resources and tools exist for assessing text readability: from this perspective, it can be said that Italian – like many other languages – is a less resourced language. In particular, we used and specialised READ–IT (Dell'Orletta et al., 2011b), the first NLP-based readability assessment tool developed for this language. READ-IT focuses on a wide range of lexical and syntactic features, whose selection was influenced by several factors, including the target application (e.g. text simplification), the intended audience as well as the intrinsic linguistic features of the language dealt with. In particular, different versions of READ-IT have been developed, implementing different approaches to readability assessment:

- classification- *vs* ranking-based approaches to readability assessment were evaluated by comparing the results of two different versions of READ-IT, namely READ-IT$_{Class}$ and READ-IT$_{Rank}$, with a focus on documents belonging to different textual genres;
- within the classification paradigm, two different versions of READ-IT$_{Class}$, called READ-IT$_{Class/Doc}$ and READ-IT$_{Class/Sent}$ respectively, were implemented with the final aim of exploring the issue of document- *vs* sentence-based readability assessment.

The paper is organized as follows: Section 2 describes the background literature on the topic, with a specific view on the open issues addressed in this paper; Sections 3 and 4 respectively illustrate the used corpora and introduce the main features underlying our approach to readability assessment;

Section 5 focuses on readability assessment through classification, which is carried out with respect to two types of textual objects, i.e. documents and sentences; finally, Section 6 tackles the issue of readability assessment across textual genres by comparing classification- and ranking-based results.

2 Background literature

Readability assessment has been a central research topic for the past 80 years which is nowadays attracting increasing attention due to the availability of advanced NLP techniques making it possible to monitor a wide variety of factors affecting the readability of a text. In what follows we will provide a survey of the literature on the topic, with particular attention to NLP-based methods and techniques.

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 et al., 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, 1988); yet, similarly to the previous case, word length does not necessarily reflect 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 et al. (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 3,000 "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 automatic readability assessment has been proposed so far, differing with respect to:

a) whether readability assessment is carried out as a classification task or in terms of ranking;
b) the typology of features taken into account (e.g. lexical, syntactic, semantic, discourse), and, within each type, the inventory of individual features used;
c) the intended audience of the text under evaluation, which strongly influences the readability assessment;
d) the application within which readability assessment is carried out.

Classification-based methods carry out readability assessment by assigning a given document to predefined readability classes. This is the approach followed in most part of the cases: see, among others, Petersen and Ostendorf (2009), Aluisio et al. (2010), Feng et al. (2010), Nenkova et al. (2010). Ranking-based approaches, positioning the document being analysed within a readability ranking scale, emerged as an alternative to classification-based methods, for dealing with less resourced languages or to meet specific needs, e.g. identifying finer-grained and customised redability classes. This class of methods is proposed, among others, by Inui and Yamamoto (2001), Tanaka-Ishii et al. (2010), Ma et al. (2012).

For what concerns the typology of features, interesting alternatives to static vocabulary-based measures such as the as the Dale-Chall formula 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 from 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 as well as the discourse cohesion and coherence, playing a central role in determining the reading difficulty of a text. More recent approaches explored the role of these features in readability assessment: this is the case, for instance, of Barzilay and Lapata (2008), Feng et al. (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 et al., 2010; Tonelli et al., 2012). Vajjala and Meurers (2012) extended further the range of potential useful features demonstrating that features inspired by Second Language Acquisition research can also contribute to improve readability assessment results.

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 target audience. Among the studies addressing readability with respect to specific audiences, it is worth mentioning here: Schwarm and Ostendorf (2005), Heilman et al. (2007), François and Fairon (2012) and Beinborn et al. (2012) all dealing with language learners, or Feng et al. (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 et. al. (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 appears to be played by discourse features (Pitler and Nenkova, 2008).

Last but not least, within the set of issues tackled in the most recent literature on readability there is the relationship between readability and textual genres: Kate et al. (2010) and Štajner et al. (2012) proved that readability features can be genre-dependent. Very recently, Sheehan et al. (2013) empirically demonstrated that, when genre effects are ignored, readability scores for informational texts (e.g. newspaper texts) tend to be overestimated, while those for literary texts (e.g. short stories, novels) tend to be underestimated: they thus proposed a two-stage approach to successfully address this problem, where the first step is aimed at identifying the text class and the second one at evaluating its readability on the basis of class-specific models. For the Italian language, Dell'Orletta et al. (2012) demonstrated that classification-based approaches to readability assessment achieve much better results by using genre-specific models, thus proving that textual genre and readability do not represent orthogonal dimensions of classification, but intertwined notions: as an alternative to generating genre-specific models, they proposed a ranking-based method for reliably assessing readability across different textual genres.

Both human- and machine-oriented applications can benefit from an automatic readability assessment. They 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. The application making use of the readability assessment, which is in turn strictly related to the intended audience of readers, can strongly influence the typology of features to be taken into account as well as the approach adopted for assessing readability (classification *vs* ranking). For instance, Beinborn et al. (2012) demonstrate that for the purposes of self-directed language learning fine-grained and personalised readability measures should be preferred over general-purpose readability predictions.

The literature reported so far focuses on readability assessment at the document level: in spite of the acknowledged need of performing readability assessment also at the sentence level (Skory and Eskenazi, 2010), it appears that sentence-level readability assessment metrics are still lacking. Interestingly, however, sentence-based analyses are reported in a text simplification scenario by De Belder and Moens (2010), Woodsend and Lapata (2011) and Drndarević et al. (2013), or for predicting the 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 the formal or informal style of a document/sentence.

Advanced NLP-based readability metrics developed so far typically deal with English, with a few attempts devoted to other languages: this is the case, for instance, of French (Collins-Thompson and Callan, 2004; François and Fairon, 2012), Portuguese (Aluisio et al., 2010), German (Brück et al., 2008; Hancke et al., 2012), Swedish (Sjöholm, 2012; Falkenjack et al., 2013) and Italian (Dell'Orletta et al., 2011b; Tonelli et al., 2012). In all cases, classification-based methods are resorted to, relying on a predefined set of readability classes (which vary according to the resources available for each language). Developed methods and techniques are mostly exploited in the framework of educational applications influencing the typology of features taken into account: consider, for instance, François and Fairon (2012) who used specific predictors of readability affecting French as L2 or Aluisio et al. (2010) who include simplification oriented features within an authoring tool for poor literacy readers. The intrinsic characteristics of the language dealt with are overtly taken into account e.g. by: Aluisio et al. (2010) and Tonelli et al. (2012), who both adapted Coh-Metrix indices for Portuguese and Italian respectively; and Dell'Orletta et al. (2011b) and Hancke et al. (2012) who both account for the effect of morphological features as valuable readability indicators for morphologically rich languages such as Italian and German. Interestingly, only Dell'Orletta et al. (2011b) and Sjöholm (2012) address the issue of assessing the readability of individual sentences as an essential prerequisite for text simplification.

3 Corpora

In assessing readability for less resourced languages, the challenge is 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, 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. As a matter of fact, the two corpora show a similar behaviour with respect to a number of different parameters, which according to the literature on register variation (Biber and Conrad, 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, they differ significantly with respect to the distribution of features typically correlated with text complexity (described in detail in Section 4 below), e.g. the composition of the used vocabulary (expressed in terms of the percentage of words belonging to the *Basic Italian Vocabulary* which 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 Sections 5 and 6. However, if on the one hand such a distinction is reliable in a document classification scenario, on the other hand at the sentence classification level it poses the remarkable issue of discerning easy–to–read sentences within difficult–to–read documents (i.e. *Rep*): 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 sentences. The implications of this fact, especially at the level of the evaluation of achieved results, are discussed in detail in Section 5.3.

4 Features

Different factors contributed to the selection of the features to be used for assessing the readability of texts. Following most NLP-based approaches described in Section 2, we consider both lexical and syntactic complexity features. Moreover, following Roark et al. (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) we included within the set of selected syntactic features also simplification oriented ones.

The features used for predicting readability are described below, 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, morpho-syntactic 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.

4.1 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.

4.2 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

DeMauro (2000), including a list of 7,000 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 out-of-vocabulary features used e.g. by Petersen and Ostendorf (2009) or François and Fairon (2012).

Lexical Variety: as stated in Bowers (2000), word repetition may affect the readability of a text. This dimension of variation of a text is typically monitored by computing the Type/Token Ratio (TTR), referring to the ratio between the number of lexical types and the number of tokens within a text. This feature has already been used for readability assessment purposes by e.g. Aluisio et al. (2010) and François and Fairon (2012). Due to its sensitivity to sample size, this feature is computed for text samples of equivalent length: in this study, TTR is calculated with respect to the first 100 tokens of a text.

4.3 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 is a quite general feature subsuming different aspects affecting in one way or another the readability of a text, ranging from POS ratios which according to Bormuth (1966) are reliable predictors as far as readability is concerned to semantically oriented features such as the "personalization level" inferred through the distribution of personal pronouns (François and Fairon, 2012). This type of feature has already been used as a reliable indicator for automatic readability assessment by, for example, Pitler and Nenkova (2008) or 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 e.g. Aluisio et al. (2010) and Feng et al. (2010).

Verbal mood: following Carreiras et al. (1997), verb tense and verb aspect appear to play an important role in the construction of mental models while reading. This feature refers to the distribution of verbs by tense and/or by mood. It is a novel and language–specific feature exploiting the predictive power of the Italian rich verbal morphology: this feature was also used by François and Fairon (2012) for the French language.

4.4 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 et al. (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 et al., 2010); the latter two are reminiscent of the 'head noun modifiers' feature used by Nenkova et al., (2010).

Verbal predicates features: this set of features captures different aspects of the behaviour of verbal predicates. They include: the *number of verbal roots* with respect to number of all sentence roots occurring in a text; the *arity of verbal predicates*, calculated as the number of instantiated dependency links sharing the same verbal head (covering both arguments and modifiers); the *distribution of verbal predicates by arity*. The relevance of this feature for readability assessment purposes is demonstrated by Kintsch et al. (1975) who showed that the number of propositions as well as the number of different arguments in a sentence influence its reading time. Dell'Orletta et al. (2011b) first used this feature for readability assessment purposes, followed more recently by François and Fairon (2012) and Falkenjack et al. (2013).

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 also with a view to the text simplification task. They include: the distribution of subordinate *vs* main clauses; the relative ordering of subordinates 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); the depth of embedded subordinate clauses measured in terms of 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. Although from a different perspective, this is also demonstrated by 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 can be seen as the dependency-based counterpart of the "phrase length" feature used for readability assessment by Nenkova et al. (2010) and Feng et al. (2010).

5 Assessing Readability through Classification

In this section, we report experiments and results achieved by the READ–IT classifier described in Section 5.1. In particular, readability classification is performed with respect to two different reading objects, namely documents (Section 5.2) and sentences (Section 5.3).

5.1 READ-IT$_{Class}$

READ–IT$_{Class}$ is a software prototype performing readability assessment of Italian texts which operates on syntactically (i.e. dependency) parsed texts and assigns to each considered reading object - either a document or a sentence - a score quantifying its readability. READ–IT$_{Class}$ 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 slightly differs from the those used at the sentence level. This also creates the prerequisites for customising the readability assessment measure with respect to target audiences characterised by specific requirements. The complete list of features used in the reported experiments was described in Section 4.

READ–IT$_{Class}$ 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.[2] Three different sets of experiments were devised to test the performance of READ-IT$_{Class}$ in the following subtasks: i) document readability classification (illustrated in Section 5.2), ii) sentence readability classification and iii) detection of easy–to–read sentences within difficult–to–read texts (both reported in Section 5.3).

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

[2] Both the Part–Of–Speech tagger and the dependency–parser represent state-of-the-art tools for the Italian language: the former shows an accuracy of 96.34% in the simultaneous identification of the grammatical category and associated morpho-syntactic features; the dependency parser shows LAS (Labelled Attachment Score) and UAS (Unlabelled Attachment Score) scores of 83.38% and 87.71% respectively.

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.

We consider the *Base Model* as our baseline: this can be seen as an approximation of the GulpEase index (i.e. the most used traditional readability measure for Italian; Lucisano and Piemontese, 1988), which is based on the same raw text features (i.e. sentence and word length).

5.2 Assessing Document Readability: Experiments and Results

For what concerns document classification, 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.

The performance of document classification experiments has been evaluated in terms of i) overall Accuracy of the system and ii) Precision, Recall and F–measure. In particular, Accuracy is a global score referring to the percentage of correctly classified documents, either as easy–to–read or difficult–to–read objects. Precision, Recall and F–measure have been computed with respect to the two target reading levels: i.e. Precision is the ratio of the number of correctly classified documents over the total number of documents classified by READ–IT$_{Class/Doc}$ 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 over the total number of documents belonging to each reading level in the test sets; F–measure is the weighted harmonic mean of Precision and Recall. For each set of experiments, evaluation was carried out with respect to the four models of the classifier.

In Table 1, the Accuracy, Precision, Recall and F–measure achieved with the different READ–IT$_{Class/Doc}$ 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 the other ones. Interestingly, the *Lexical Model* shows a high accuracy (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-syntactic information.

		2Par			*Rep*		
Model	Accuracy	Prec	Rec	F-measure	Prec	Rec	F-measure
Base	76.65	74.71	80.56	77.52	78.91	72.73	75.69
Lexical	95.45	95.60	95.30	95.45	95.31	95.61	95.46
MorphoS	**98.12**	**98.12**	**98.12**	**98.12**	**98.12**	**98.12**	**98.12**
Syntax	97.02	97.17	96.87	97.02	96.88	97.18	97.03

Table 1: Document classification results by the different models

5.3 Assessing Sentence Readability: Experiments and Results

With regard to the sentence classification subtask, we used a training set of randomly selected documents corresponding to about 3,000 sentences from *2Par* and about 3,000 sentences from *Rep*. Two different experiments were performed operating on test sets of different size. A first experiment was carried out on 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*). A second experiment was carried out 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.

Different evaluation methods have been defined in order to assess achieved results in the two experiments focusing on sentence readability assessment. As already done with documents, the performance of sentence classification experiments has first been evaluated in terms of i) overall Accuracy of the system and ii) Precision, Recall and F-measure. However, this type of evaluation might not be so appropriate in the case of sentences: 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. The lack of training data at the sentence level makes it difficult – if not impossible – to evaluate the effectiveness of the proposed 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.

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$_{Class/Sent}$ 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 to discern errors from "correct" misclassifications, we introduced a new evaluation methodology, based on the notion of Euclidean distance between n–dimensional vectors of feature–values (for the typology of features see Section 4). Each feature vector represents a set of sentences and is obtained by averaging the vectors corresponding to the individual sentences constituting the set. Two vectors with 0 distance represent sets of 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. This notion of distance was also used to test which model was more effective in predicting the readability of n–word long sentences.

Consider now the first experiment. Table 2 shows that the most reliable results at the sentence level 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, Recall and F-measure 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 (where the *Syntax Model* is the best performing one) is much higher, namely 17%.

		2Par			*Rep*		
Model	Accuracy	Prec	Rec	F-measure	Prec	Rec	F-measure
Base	59.60	55.62	95.00	70.16	82.88	24.20	37.46
Lexical	61.60	57.30	91.00	70.32	78.16	32.20	45.61
MorphoS	76.10	72.78	83.40	77.72	80.56	68.80	74.22
Syntax	**78.20**	**75.09**	**84.40**	**79.47**	**82.19**	**72.00**	**76.76**

Table 2: Sentence classification results by the different models

Table 3 refers to the performance of the best READ–IT$_{Class/Sent}$ 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 (360 sentences, referred to as *correct* in the Table), iii) the whole *Rep test set*. It can be noticed that the Euclidean distance between the *wrong* sentences and *2Par* is much lower than the distance holding between *2Par* and the correcly classified sentences (*correct*). This result suggests 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*). As expected, the *Rep test set* as a whole, containing both *correctly* and *wrongly* classified sentences, has an intermediate distance value with respect to *2Par*.

The reliability of this distance-based method was assessed by manually performing a qualitative analysis of the whole *Rep test set*, i.e. the 140 *Rep* sentences erroneously classified as belonging to *2Par* and the 360 correctly classified *Rep* sentences. This analysis was carried out by two human annotators who were asked to independently judge sentences as easy–to–read or difficult–to–read according to their own perception and intuition: the annotators, native speakers of Italian, were not provided with explicit annotation guidelines to discern between the two cases but were trained on the basis of examples of easy–to–read sentences. In the manual evaluation of the set of 140 erroneously classified sentences, the first annotator classified 107 sentences as easy–to–read and the remaning ones (33) as difficult–to–read, whereas the second annotator identified 120 easy–to–read sentences against 20 complex ones. The proportion of sentences on which the two annotators turned out to agree is high, namely 78% corresponding to 99 sentences classified by both as easy–to–read and 10 classified as difficult-to-read. For what concerns the 360 correctly classified *Rep* sentences, the first annotator classified 330 sentences as difficult–to–read and 30 sentences as easy–to–read, whereas the second annotator identified 318 difficult–to–read sentences against 42 easy–to–read: it is interesting to note that the two annotators did not agree in the classification of only 12 sentences, showing a much higher level of agreement with respect to the previous case. Taken together, the annotators agreed in 91,4% of the cases, with a significant difference between the two subsets of the *Rep test set*: the lower agreement observed on the 140 sentences erroneously classified as belonging to *2Par* demonstrates that this represents a set of sentences that cannot be clearly classified as easy– or difficult–to–read. The results of this manual analysis are in line with those automatically computed on the basis of the distance between feature vectors, thus demonstrating the reliability of the adopted distance-based evaluation methodology.

	Distance
Correct	52.07
Rep test set	45.36
Wrong	37.84

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

	Accuracy
Base	23.39
Lexical	38.71
MorphoS	70.49
Syntax	**76.01**

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

The second experiment operates on a much wider test set, i.e. *Rep 2.5*. Table 4 reports the accuracy achieved by the different models in the classification of difficult–to–read sentences. According to the reported results, it appears that the *Syntax Model* classifies the higher number of difficult–to–read sentences: since *Rep 2.5* sentences are not annotated with readability information, this result alone is not sufficient to prove whether and to what extent this is the best performing model. In order to compare the performance of the four READ–IT models and to identify the most effective one, we followed the distance-based evaluation methodology described above: i.e. we computed the Euclidean distance between n–dimensional vectors of feature–values representing a given set of sentences (e.g. those classified as easy-to-read by a specific model) and *2Par*. Note that each set of sentences is represented by a vector containing all features described in Section 4. Table 5 reports, for each model, the results of this distance-based evaluation. If on the one hand it can be noticed that easy–to–read sentences according to the *Syntax Model* show the lowest distance with respect to *2Par*, on the other hand the whole *Rep 2.5* corpus shows a higher distance due to the fact it includes both difficult– and easy–to–read sentences; it goes without saying that the sentences classified as difficult–to–read by the *Syntax Model* (*Diff Syntax* in the Table) show the largest distance. The accuracy achieved in the classification of difficult–to–read sentences combined with the shortest distance of easy–to–read sentences with respect to *2Par* suggests that the *Syntax Model* is the most effective one as far as sentence readability assessment is concerned.

	Distance
Diff Syntax	66.53
Rep 2.5	64.04
Base	61.14
Lexical	60.53
MorphoS	55.54
Syntax	**51.41**

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.

Further evidence in this direction was gained by comparing the performance of the four READ–IT$_{Class/Sent}$ models for sets of sentences of fixed length. In particular, we considered sentences ranging in length from 8 to 30 words. For every set of sentences of equivalent length, we compared the *Rep 2.5* sentences classified as easy–to-read by the four models with respect to *2Par*. In Figure 1, each point represents the Euclidean distance between the feature vectors representing a set of sentences of a given length and the same n–word long set of sentences in the *2Par* corpus (see Section 4 for the typology of features in each vector). As it can be seen, the bottom line, which represents the set of sentences classified as easy–to–read by the *Syntax Model*, is the closest to the *2Par* sentences of equivalent length. On the contrary, the line representing the sentences classified as easy–to-read by the *Base Model* is the most distant amongst the four READ–IT$_{Class/Sent}$ models. Interestingly, it overlaps with the line representing the whole set of *Rep 2.5* sentences: this suggests that a classification model based on raw text features (i.e. sentence and word length) only is not able to reliably identify easy–to–read sentences if we consider sets of sentences of a fixed length. Obviously, the line representing the set of sentences classified as difficult–to–read by the *Syntax Model* shows the largest distance. In this second experiment, we demonstrated that linguistically motivated features (and in particular syntactic ones) play a fundamental role in sentence readability assessment.

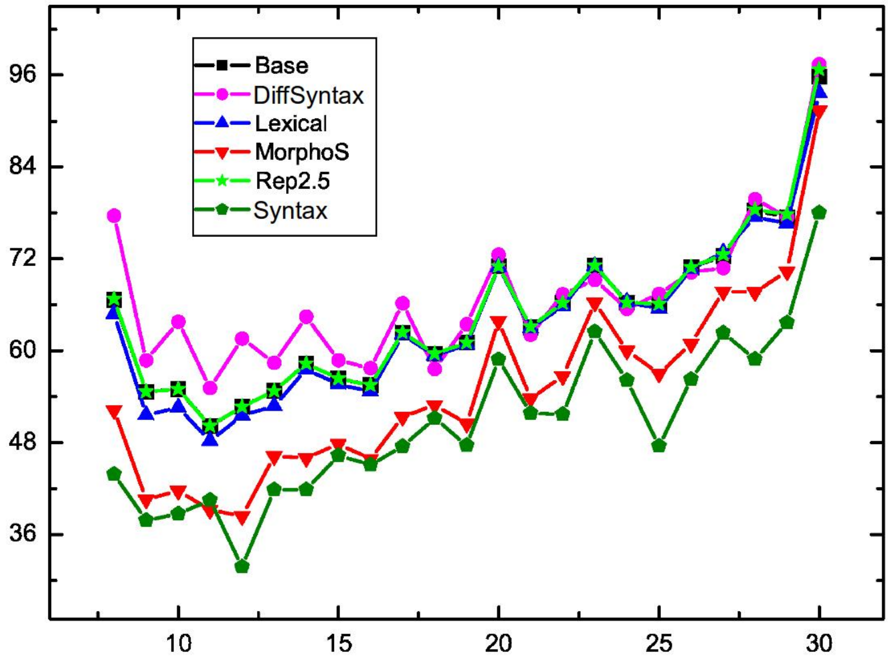

Figure 1: Distance, for sets of sentences of fixed length, between *2Par* and i) difficult–to–read sentences according to the *Syntax Model*, ii) *Rep 2.5*, iii) easy–to–read sentences by the four models

6 Readability Assessment and Textual Genres

In this section, we explore the complex interplay between genre and readability analysis, with the final aim of identifying workable solutions which might be exploited in real–world, e.g. educational, applications. This goal was pursued in two different steps. Firstly, we demonstrate that readability assessment is genre–dependent, by carrying out two classification–based readability assessment experiments using a single readability model *vs* genre–specific models and by comparing results achieved in classifying documents which belong to different genres. Secondly, we propose a new ranking–based readability assessment method exploiting complex NLP–enabled linguistic features.

6.1 Corpora

To explore the relationship holding between genre and readability analysis, we focused on four traditional textual genres: Journalism, Literature, Educational writing and Scientific prose. For each genre, the corpus of texts was further subdivided in two classes according to their expected target audience, taken as indicative of the accessibility level of the document. For the journalistic genre, we used the *Rep* and *2Par* corpora described in Section 3. The corpora selected as representative of the Literature and Educational genre classes were partitioned into two sub-corpora, including texts respectively targeting children *vs* adults. The scientific prose genre class includes articles from Wikipedia and scientific literature, which are taken as representative, respectively, of the easy–to–read and difficult–to–read classes within the scientific prose genre: the choice of taking Wikipedia articles as easy–to–read documents was based on the wide audience they target. Among all these corpora, due to its peculiar nature (i.e. the fact of resulting from a simplification process) *2Par* is to be considered as the easiest–to–read corpus. Corpora selected as representative of the different genre classes and accessibility levels are detailed in Table 6.

Abbreviation name	**Corpus**	**Coarse–grained genre**	**N. documents**	**N. words**
Rep	*La Repubblica*, Italian newspaper Marinelli et al. (2003)	*Journalism*	321	232,908
2Par	*Due Parole*, easy–to–read Italian newspaper Piemontese (1996)	*Journalism*	322	73,314
ChildLit	*Children Literature* Marconi et al. (1994)	*Literature*	101	19,370
AdLit	*Adult Literature* Marinelli et al. (2003)	*Literature*	327	471,421
ChildEdu	*Educational Materials* for Primary School Dell'Orletta et al. (2011a)	*Educational*	127	48,036
AdEdu	*Educational Materials* for High School Dell'Orletta et al. (2011a)	*Educational*	70	48,103
Wiki	*Wikipedia* articles from the Italian Portal "Ecology and Environment"	*Scientific prose*	293	205,071
ScientArt	*Scientific articles* on different topics (e.g. climate changes and linguistics)	*Scientific prose*	84	471,969

Table 6: Corpora representative of different textual genres

It is interesting to note that besides Wikipedia, which focuses on environmental topics, all other corpora are not restricted to a given topic but rather cover a variety of them. This is the case of newspaper articles dealing with different topics, of educational materials covering different subjects, of the scientific literature (on climate change and linguistics) as well as of the literature genre represented by different types of narrative prose. This was done to neutralise, or at least to reduce, influences possibly deriving from the topic–genre correlation reported in the literature (see, e.g., Petrenz and Webber, 2011).

For the experiments reported below, each corpus representative of a fine–grained subclass, corresponding to a textual genre and targeting a specific audience, was split into training and test sets. Each test set consists of 30 selected documents, whereas the training sets include the remaining documents, namely: 292 (*2Par*), 291 (*Rep*), 71 (*ChildLit*), 297 (*AdLit*), 97 (*ChildEdu*), 40 (*AdEdu*), 263 (*Wiki*) and 54 (*ScientArt*).

6.2 Readability Classification Across Textual Genres

In order to explore whether and to what extent readability is related to textual genre, we carried out two sets of experiments aimed at discerning, within each of the four genre classes, easy– vs difficult–to–read documents and differing at the level of used models: in the first set, we used a single and general statistical model for all four genres, whereas in the second set readability classification was performed by using genre–specific statistical models. Achieved results have been evaluated in terms of i) overall Accuracy of the system and ii) Precision, Recall and F–measure.

In the first set of experiments, we tested three models differing at the level of the used training sets. For the first model, the training corpora for easy– and difficult–to–read documents are represented by newspaper texts, i.e. belonging to the same genre: as discussed in Dell’Orletta et al. (2011b), this prevents interferences due to textual genre variation in the measure of text readability. For the second model, documents belonging to two different genres were selected for training: i.e. *2Par* was used as representative of the easy–to–read class, whereas for the difficult–to–read class we chose the *ScientArt* corpus. This option followed from the fact that the newspaper articles of *2Par* represent the most accessible documents in the collection we have been dealing with, while the scientific articles included in the *ScientArt* corpus turned out to be the most difficult ones (see Section 6.3). For the third model, the training sets have been constructed by combining all the easy–to–read and difficult–to–read documents for each textual genre respectively. In Table 7, the columns headed by *2Par/Rep Model*, *2Par/ScientArt Model* and *All Easy/All Difficult Model* show the results achieved for each textual genre with the three models just described. In the last set of rows, Precision, Recall, F–measure and Accuracy scores for the whole set of documents (i.e. regardless of genre) are reported. The *2Par/Rep Model* turned out to obtain the best results. However, none of the three models achieves noteworthy results when compared with those obtained in the document readability classification task reported in Section 5.2 (i.e. 98.12%). This suggests that classification–based methods are able to assign a reliable readability score only when dealing with documents belonging to the same genre as the training set: see the Accuracy obtained by the *2Par/Rep Model* tested on texts of the same journalistic genre (98.33%). In all other cases, the results achieved show that this method has a dramatic drop in accuracy when tested on documents belonging to different genres with respect to the training set.

Consider now the results of the second set of experiments carried out using a specific model for each of the four genres, reported in Table 8. As expected, the overall accuracies significantly increase with respect to the results obtained by the general models. The only exception is represented by the classification of the documents in the class of *Scientific writing* characterised by a much lower Accuracy, with a Recall of 13.33% obtained in the *ScientArt* document classification and a Precision of 53.57% in the *Wiki* classification. We can hypothesize that this result follows from the internal composition of the *Wiki* training set, which might not include easy–to-read documents only.

	2Par/Rep Model			2Par/ScientArt Model			All Easy/All Difficult Model		
Genre	Prec	Rec	F–measure	Prec	Rec	F–measure	Prec	Rec	F–measure
2Par	100.00	96.67	98.30	50.85	100.00	67.41	93.55	96.67	95.08
Rep	96.78	100.00	98.36	100.00	3.33	6.45	96.55	93.33	94.91
	Accuracy: 98.33			**Accuracy: 51.67**			**Accuracy: 95**		
ChildLit	0	0	0	46.81	73.33	57.14	100.00	46.67	63.63
AdLit	50	100.00	66.67	38.46	16.67	23.25	65.22	100.00	78.95
	Accuracy: 50			**Accuracy: 45**			**Accuracy: 73.33**		
ChildEdu	90	31.03	46.15	49.15	100.00	65.91	56.67	58.62	57.63
AdEdu	59.18	96.67	73.42	0	0	0	58.62	56.67	57.63
	Accuracy: 64.41			**Accuracy: 49.15**			**Accuracy: 57.63**		
Wiki	100.00	20	33.33	81.25	86.67	83.87	47.17	83.33	60.24
ScientArt	55.55	100.00	71.43	85.71	80	82.76	28.57	6.67	10.81
	Accuracy: 60			**Accuracy: 83.33**			**Accuracy: 45**		
TOT Easy–to–read	97.78	36.97	53.66	54.31	89.91	67.72	66.40	71.43	68.82

TOT Difficult–to–read	61.34	99.17	75.80	71.43	25	37.04	69.34	64.17	66.67
	Accuracy: 68.20			**Accuracy: 57.32**			**Accuracy: 67.78**		

Table 7: Classification–based readability assessment results by general models

	Genre–specific Models		
Genre	Prec	Rec	F–measure
2Par	100.00	96.67	98.30
Rep	96.77	100.00	98.36
	Accuracy: 98.33		
ChildLit	84.61	73.33	78.57
AdLit	76.47	86.67	81.25
	Accuracy: 80		
ChildEdu	78.79	89.65	83.87
AdEdu	88.46	76.67	82.14
	Accuracy: 83.05		
Wiki	53.57	100.00	69.77
ScientArt	100.00	13.33	23.53
	Accuracy: 56.67		
TOT Easy–to–read	74.30	89.91	81.37
TOT Difficult–to–read	87.37	69.17	77.21
	Accuracy: 79.51		

Table 8: Classification–based readability assessment results by genre-specific models

The results of the experiments reported so far show that readability assessment is closely related to the textual genre of a document, suggesting that for reliably dealing with different textual genres a specific training corpus should be built for each genre. This represents an unrealistic goal to be pursued, especially in real–world applications. In what follows, an alternative approach to the problem is presented, i.e. a ranking method able to reliably assign a readability score without requiring genre–specific training corpora.

6.3 Assessing Readability Across Genres by Ranking

Our ranking–based approach to readability assessment is grounded on the notion of cosine distance between vectors of linguistic features. The readability score is computed as a linear combination between the distance of the vector representing an analysed document (*d*) and two n–dimensional vectors corresponding to the easy (*EV*) and the difficult–to–read poles (*DV*):

$$readability(d) = CosineDistance(d,EV) - CosineDistance(d,DV)$$

According to the equation, the readability score ranges from −1 (easy–to–read document) to 1 (difficult–to-read document). To cope with the fact that the distance from, e.g., the easy extreme (*EV*) can express the difficulty but also the extreme readability of *d*, in the final score we combined the distance from both *EV* and *DV* poles. With respect to the ranking method proposed by Tanaka-Ishii et al. (2010), we assign to each analyzed document a score rather than a relative ranking position, making less questionable the comprehension of the results. From the computational point of view, our method, based on the notion of *distance*, is much less complex than the ranking method by Tanaka-Ishii et al. (2010) which is based on a *comparison* strategy.

As stated in Section 6.1, we assumed *2Par* as the easiest–to–read pole. The difficult–to-read extreme was identified by computing the cosine distance between the *2Par* vector and the feature vectors representing the remaining sub-corpora (resulting from the combination of textual genre and target audience): as already stated in Section 5.3, each feature vector representing a set of sentences is obtained by averaging the vectors corresponding to the individual sentences constituting the set. The *ScientArt* vector turned out to be the most distant one from *2Par* and for this reason it was selected as the difficult extreme. We report below the ordered list of the eight fine–grained subclasses considered in this study ranked by increasing distance from *2Par*:

2Par < ChildEdu < ChildLit < Rep < Wiki < AdLit < AdEdu < ScientArt

This ranking was computed twice, with respect to the vectors representing i) the training sets (bigger but of unequal size) and ii) the much smaller test sets (all of the same size): in both bases, the same result was obtained. Note that for each genre the relative ordering of easy– *vs* difficult–to–read subclasses is preserved. It is also worth noting the ranking of *Rep* before *Wiki* which can be taken as further evidence of the difficulty of defining a readability notion valid across all genres.

In Table 9, the list of test documents ranked by increasing distance-based readability scores is grouped into sets of 30 documents, where the top document classes are assumed to be easier to read than the bottom classes. Each row represents the distribution across genres of the represented set of 30 documents. Interestingly, for each genre the number of easier to read documents, i.e. closer to *2Par*, is higher in the top 30–document groups whereas the reverse holds in the bottom. However, the distribution of easy– *vs* difficult–to–read documents is not homogeneous across genres: whereas for

2Par, *ChildLit* and *ChildEdu* the distribution across the 30–document groups meets the expectations, *Wiki* documents appear to be homogeneously distributed in all classes. For what concerns difficult–to–read documents, it can be noticed that *AdLit*, *AdEdu* and *ScientArt* documents tend to concentrate in the mid and bottom groups; the only exception is represented by *Rep* documents which are distributed in all classes except the top one, with higher frequencies in the bottom 30-document groups.

	Journalism		Literature		Educational		Scientific prose	
Doc.Group	2Par	Rep	ChildLit	AdLit	ChildEdu	AdEdu	Wiki	ScientArt
0-30	15	0	4	0	8	0	3	0
31-60	6	1	11	0	9	0	3	0
61-90	4	6	7	6	3	1	1	0
91-120	1	5	1	12	2	5	4	0
121-150	2	3	2	7	5	6	4	1
151-180	1	1	2	3	2	11	4	6
181-210	1	8	2	2	1	3	5	8
211-240	0	6	1	0	0	4	4	15

Table 9: Ranking–based readability assessment results

7 Conclusion

In this paper, we tackled three open issues of the automatic readability assessment literature, namely: the evaluation of text readability in a language like Italian which is low-resourced as far as readability is concerned, with respect to sentences (as opposed to documents) as well as across textual genres. Different solutions to the issues addressed have been tested on Italian texts by using and specialising READ–IT, the first advanced readability assessment tool for this language. READ-IT focuses on a wide range of raw, lexical, morpho-syntactic and syntactic features, which were selected with a view to potential target applications which could usefully exploit the readability assessment results (e.g. text simplification), the intended audience as well as intrinsic linguistic features of Italian (e.g. being a morphologically rich language). Different versions of READ-IT were developed, implementing different approaches to readability evaluation.

The issue of document- *vs* sentence-based readability assessment was explored within the classification paradigm, by comparing the results of READ-IT$_{\mathit{Class/Doc}}$ and READ-IT$_{\mathit{Class/Sent}}$: sentence–based readability assessment is an important novelty of our approach which has the further bonus of creating 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 tasks appear to enforce different requirements at the level of the underlying linguistic features (the *MorphoS Model* is the best performing one for what concerns document classification, whereas the *Syntax Model* achieves the best results as fas as sentence readability is concerned). To overcome the lack of an Italian reference corpus annotated with readability information at the sentence level, the notion of distance was resorted to to assess the performance of READ–IT; reliability and effectiveness of this notion were validated through a manual revision of the whole test set.

Advantages and drawbacks of classification- *vs* ranking-based approaches to readability assessment were explored by comparing the results of READ-IT$_{Class}$ and READ-IT$_{Rank}$ in the analysis of documents representative of four traditional textual genres. We have shown that readability assessment is strongly influenced by textual genre: with classification–based approaches reliable results can only be achieved with genre–specific models. Since this is far from being a workable solution, especially for less–resourced languages, we proposed a new ranking method for readability assessment based on the notion of distance. The potential impact of the newly introduced ranking–based readability assessment method was explored in the automatic construction of genre–specific training sets which we have seen to be deeply needed to achieve reliable results in classification–based readability assessment. To this end, a pilot experiment was performed focusing on the *Scientific writing* genre for which the most unsatisfactory results were reported. To improve the accuracy of the classification within this class, we automatically revised the *Wiki* training set using the newly proposed distance readability score with the aim of selecting easy–to-read documents only. In particular, we ranked the documents contained in the original *Wiki* training set and picked the top list of 100 documents, which was used as the new training set. Table 10 reports the results of READ–IT$_{Class/Doc}$ with the new genre–specific model built using the automatically constructed *Wiki* training set, where an improvement of 21.66% in accuracy is observed with respect to the previous results recorded in Table 8: this demonstrates the reliability of the proposed ranking method, as well as its usefulness in the construction and/or refinement of training resources.

Genre	Prec	Rec	F–measure
Wiki	72.97	90.00	80.60
ScientArt	86.96	66.67	75.47
	Accuracy: 78.33		

Table 10: Classification results on *Scientific prose* using the automatically revised training set

8 References

Aluisio, S., Specia, L., Gasperin, C. & Scarton, C. (2010). Readability assessment for text simplification. In J. Tetreault, J. Burstein & C. Leacock (Eds.), *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 1–9). Los Angeles, California: Association for Computational Linguistics.

Attardi, G. (2006). Experiments with a multilanguage non-projective dependency parser. In L. Màrquez & D. Klein (Eds.), *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X '06)* (pp. 166–170). New York City: Association for Computational Linguistics, http://www.aclweb.org/anthology/W/W06/W06-13.

Barzilay, R. & Lapata, M. (2008). Modeling local coherence: An entity-based approach. *Computational Linguistics, 34(1)*, 1–34.

Beinborn, L., Zesch, T. & Gurevych, I. (2012). Towards fine-grained readability measures for self-directed language learning. In *Proceedings of the SLTC 2012 Workshop on NLP for CALL*, Volume 2 (pp. 11-19). Lund (Sweden): Öping University Electronic Press.

Biber, D. & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Bormuth, J. R. (1966). Readability: A new approach. *Reading research quarterly*, *1*, 79-132.

Bowers, J.S. (2000). In defense of abstractionist theories of repetition priming and word identification. *Psychonomic Bulletin & Review*, *7*, 83-99.

Caldwell, B., Cooper, M., Guarino Reid, L. & Vanderheiden. G. (Eds.) (2008). *Web Content Accessibility Guidelines 2.0*. World Wide Web Consortium, Recommendation REC-WCAG20-20081211, (December 2008), http://www.w3.org/TR/WCAG20/.

Carreiras, M., Carriedo, N., Alonso, M.A. & Fernández, A. (1997). The role of verb tense and verb aspect in the foregrounding of information during reading. *Memory & Cognition*, *25(4)*, 438–446.

Chall, J.S. & Dale, E. (1995). *Readability Revisited: The New Dale–Chall Readability Formula*. Cambridge, MA: Brookline Books.

Chang, C-C. & Lin, C-J. (2001). *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Collins-Thompson, K. & Callan, J. (2004). A language modeling approach to predicting reading difficulty. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004)* (pp. 193–200). Boston, Massachusetts, USA: Association for Computational Linguistics.

De Belder, J., & Moens, M-F. (2010). Text simplification for children. *Proceedings of the SIGIR Workshop on Accessible Search Systems*, 19-26. New York: ACM.

De Mauro, T. (2000). *Il dizionario della lingua italiana*. Paravia: Torino.

Dell'Orletta, F. (2009). Ensemble system for Part-of-Speech tagging. *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence,* 12th December 2009, Reggio Emilia, Italy, ISBN 978-88-903581-1-1.

Dell'Orletta, F., Montemagni, S. & Venturi, G. (2011b). Read-it: Assessing readability of italian texts with a view to text simplification. In N. Alm (Ed.), *Proceedings of the Second Workshop on "Speech and Language Processing for Assistive Technologies" (SLPAT 2011),* 30 July 2011, Edinburgh, UK (pp. 73-83). Edinburgh, Scotland, UK: Association for Computational Linguistics.

Dell'Orletta, F., Montemagni, S. & Venturi, G. (2012). Genre-oriented Readability Assessment: a Case Study. In R. Mamidi & K. Prahallad (Eds.), *Proceedings of the COLING-2012 Workshop on Speech and Language Processing Tools in Education (SLP-TED),* 15 December 2012, Mumbai, India (pp. 91-98).

Dell'Orletta, F., Montemagni, S., Vecchi, E.M. & Venturi, G. (2011a). Tecnologie linguistico–computazionali per il monitoraggio della competenza linguistica italiana degli alunni stranieri nella scuola primaria e secondaria. In G.C. Bruno, I. Caruso, M. Sanna & I. Vellecco (Eds.), *Percorsi migranti: uomini, diritto, lavoro, linguaggi* (pp. 319-366). Milano: McGraw–Hill Editore.

Drndarević, B., Štajner, S., Bott, S., Bautista, S. & Saggion, H. (2013). Automatic Text Simplification in Spanish: A Comparative Evaluation of Complementing Modules. In A. Gelbukh (Ed.), *Proceedings of the Computational Linguistics and Intelligent Text Processing - 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013,* Part II (pp. 488-500). Berlin Heidelberg: Springer-Verlag, LNCS 7817.

Falkenjack, J., Mühlenbock, K.H. & Jönsson, A. (2013). Features indicating readability in Swedish text. *Proceedings of the 19th Nordic Conference of Computational Linguistics*, 27-40.

Feng, L., Elhadad, N. & Huenerfauth, M. (2009). Cognitively motivated features for readability assessment. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL '09)*, 229-237.

Feng, L., Jansche, M., Huenerfauth, M. & Elhadad, N. (2010). A comparison of features for automatic readability assessment. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010),* 276-284.

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

François, T. & Fairon, C. (2012). An "AI readability" formula for French as a foreign language. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jeju Island, Korea, 466-477.

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

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

Hancke, J., Vajjala, S. & Meurers, D. (2012). Readability Classification for German using Lexical, Syntactic, and Morphological Features. *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, Mumbai, India, 1063-1080.

Heilman, M.J., Collins, K. & Callan, J. (2007). Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts. *Proceedings of the Human Language Technology Conference*, pp. 460-467.

Inui, K. & Yamamoto, S. (2001). Corpus-based acquisition of sentence readability ranking models for deaf people. *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*, Tokyo, 159-166.

Kate, R.J., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R.J., Roukos, S. & Welty, C. (2010). Learning to Predict Readability using Diverse Linguistic Features. *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, 546-554.

Kincaid, J. P., Fishburne, L.R.P., Rogers, R.L. & Chissom, B.S. (1975). *Derivation of new readability formulas for Navy enlisted personnel* (pp. 8–75). Research Branch Report, Millington, TN: Chief of Naval Training.

Kintsch, W., Kozminsky, E., Streby, W.J., McKoon, G. & Keenan, J.M. (1975). Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior*, 14(2), 196–214.

Lin, D. (1996). On the structural complexity of natural language sentences. *Proceedings of COLING 1996*, 729-733.

Louis, A. & Nenkova, A. (2013). A corpus of science journalism for analysing writing quality. *Dialogue and Discourse*, 4(2), 87-117.

Lucisano, P. & Piemontese, M.E. (1988). GulpEase. Una formula per la predizione della difficoltà dei testi in lingua italiana. *Scuola e Città* (3), 57–68.

Ma, Y., Fosler-Lussier, E. & Lofthus, R. (2012). Ranking-based readability assessment for early primary children's literature. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Montréal, Canada, 548-552.

Marconi, L., Ott, M., Pesenti, E., Ratti, D. & Tavella, M. (1994). *Lessico Elementare*. Bologna: Zanichelli.

Marinelli, R., Biagini, L., Bindi, R., Goggi, S., Monachini, M., Orsolini, P., Picchi, E., Rossi, S., Calzolari, N. & Zampolli, A. (2003). The italian parole corpus: an overview. In Zampolli, A. et al. (Eds.), *Computational Linguistics in Pisa*, *Special Issue*, XVI–XVII, Tomo I (pp. 401–421). Pisa: IEPI.

McDonald, R. & Nivre, J. (2007). Characterizing the Errors of Data-Driven Dependency Parsing Models. *Proceedings of EMNLP-CoNLL 2007*, 122-131.

Miller, J. & Weinert, R. (1998). *Spontaneous spoken language. Syntax and discourse*. Oxford, Clarendon Press.

Nenkova, A., Chae, J., Louis, A. & Pitler, E. (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* (pp. 222–241). Berlin Heidelberg: Springer-Verlag, LNAI 5790.

Petersen, S.E. & Ostendorf, M. (2006). *A machine learning approach to reading level assessment*. University of Washington CSE Technical Report.

Petersen, S.E. & Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech and Language*, (23), 89–106.

Petrenz, P., & Webber, B. (2011). Stable classification of text genres. *Computational Linguistics* 37 (2), 385-393.

Piemontese, M.E. (1996). *Capire e farsi capire. Teorie e tecniche della scrittura controllata* Napoli, Tecnodid.

Pitler, E. & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 186-195.

Roark, B., Mitchell, M. & Hollingshead, K. (2007). Syntactic complexity measures for detecting mild cognitive impairment. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing,* 1-8.

Schwarm, S.E. & Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL 05),* 523-530.

Sheehan, K.M., Flor, M. & Napolitano, D. (2013). A Two-Stage Approach for Generating Unbiased Estimates of Text Complexity. *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, Atlanta, Georgia, 49-58.

Sheikha, F.A. & Inkpen, D. (2012). Learning to Classify Documents According to Formal and Informal Style. *Linguistic Issues in Language Technology*, 8(1), 1-29.

Si, L. & Callan, J. (2001). A statistical model for scientific readability. *Proceedings of the tenth international conference on Information and knowledge management,* 574-576.

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

Skory, A. & Eskenazi, M. (2010). Predicting cloze task quality for vocabulary training. *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications,* 49-56.

Štajner, S., Evans, R., Orasan, C. & Mitkov, R. (2012). What can readability measures really tell us about text complexity? *Proceedings of the the Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, Istanbul, Turkey, 14-21.

Tanaka-Ishii, K., Tezuka, S. & Terada, H. (2010). Sorting texts by readability. *Computational Linguistics*, 36, 2, 203–227. Cambridge, MA, USA: MIT Press.

Tonelli, S., Manh, K.T. & Pianta, E. (2012). Making readability indices readable. *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, Montréal, Canada, 40-48.

Vajjala, S. & Meurers, D. (2012). On improving the accuracy of readability classification using insights from second language acquisition. *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP,* Montréal, Canada, 163-173.

vor der Brück, T., Hartrumpf, S. & Helbig, H. (2008). A Readability Checker with Supervised Learning using Deep Syntactic and Semantic Indicators. *Proceedings of the 11th International Multiconference: Information Society - IS 2008 - Language Technologies*, Ljubljana, Slovenia, 92-97.

Woodsend, K., & Lapata, M. (2011). Learning to Simplify Sentences with Quasi-synchronous Grammar and Integer Programming. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2011)*, 409-420.

Yngve, V.H.A. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, 444-466.

Zipf, G.K. (1988). *The Psychobiology of Language*. Houghton–Miflin, Boston.