Genre–oriented Readability Assessment: a Case Study

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

ABSTRACT
Whether and to what extent readability assessment is genre–dependent is an issue which has important consequences also in the design and development of educational applications. In this paper, we address this issue from an applicative point of view by investigating whether general purpose readability assessment tools can reliably be used for dealing with texts belonging to different genres. Different experiments have been carried out showing that classification–based approaches to readability assessment can achieve reliable results only by using genre–specific models. Since the construction of genre–specific models is a time consuming task, we proposed a new ranking method for readability assessment based on the notion of distance: we also showed that this method can be usefully exploited for automatically building genre–specific training corpora, thus creating the prerequisites for overcoming the inherent problems of classification–based readability assessment. All reported experiments have been carried out on Italian, a less—resourced language as far as readability assessment is concerned.

Valutazione della Leggibilità e Generi Testuali: un Caso di Studio

Se e in che misura la valutazione della leggibilità sia influenzata dal genere testuale rappresenta una questione che ha importanti conseguenze anche al livello dello sviluppo di applicazioni in ambito didattico. In questo contributo, affrontiamo questo problema da una prospettiva applicativa, verificando se strumenti per il calcolo della leggibilità sviluppati per un uso generale siano affidabili quando applicati a testi appartenenti a diversi generi testuali. Sono stati condotti diversi esperimenti che hanno mostrato che approcci al calcolo della leggibilità basati sul metodo della classificazione possono restituire risultati affidabili solo se utilizzano modelli specifici per ogni genere. Dal momento che la costruzione di tali modelli specifici è un compito impegnativo, abbiamo proposto un nuovo metodo di ranking per il calcolo della leggibilità basato sulla nozione di distanza che può essere utilizzato anche per la costruzione automatica di corpora di addestramento specifici di genere. Tutti gli esperimenti riportati sono stati condotti sull'italiano, lingua per la quale sono a disposizione poche risorse.

KEYWORDS: Readability Assessment, Textual Genre, Automatic Construction of Corpora, Less–resourced Languages.

KEYWORDS IN L_2: Calcolo della Leggibilità, Generi Testuali, Costruzione Automatica di Corpora, Lingue con Poche Risorse.

Proceedings of the Workshop on Speech and Language Processing Tools in Education, pages 91–98,
COLING 2012, Mumbai, December 2012.

1 Introduction

Over the last ten years, the development of efficient natural language processing (NLP) systems led to a resurgence of interest in readability assessment. Several studies have been carried out based on NLP-enabled feature extraction and state–of–the–art machine learning algorithms with significant performance improvement with respect to traditional readability measures. Due to the great potential of automatic readability assessment for educational purposes, many of these studies, mostly focused on English but also tackling less–resourced languages, have been carried out with the final aim of supporting teachers and/or learners in selecting material which is appropriate to a given reading level. In principle, educational material can belong to different textual genres, ranging e.g. from fiction to scientific writing or reportage. The question which naturally arises is whether and to what extent readability assessment is genre–independent, and if this not the case whether and how general purpose readability assessment tools could reliably be used for dealing with texts belonging to different genres. The most recent literature on readability reports that the degree of readability is connected to genre: 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 educational applications.

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; Dell'Orletta et al., 2011b)) 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 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. The main problem of classification–based 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 less–resourced languages. Moreover, if it turns out to be true that the notion of readability has to be tailored with respect to textual genres, such resources would in principle be needed for each genre: this represents an unrealistic goal. Ranking–based readability assessment methods represent a viable alternative to classification methods, since they only require training data with respect to two readability levels (easy vs difficult).

The work reported in this paper is aimed at shedding light on the complex interplay between genre and readability analysis, with the goal of exploring workable solutions which might be exploited in real–world educational applications. This goal was pursued in two different steps. Firstly, we demonstrated that readability assessment is genre–dependent: we carried out two classification–based readability assessment experiments and compared the results achieved in classifying documents which belong to different genres using a single readability model and genre–specific models. Secondly, we proposed a new ranking–based readability assessment method exploiting complex linguistic features identified within the output of NLP tools. All reported experiments have been carried out on a less–resourced language, i.e. Italian.

2 Corpora and Tools

For the specific concerns of this study, we focused on four traditional textual genres: Journalism, Literature, Educational writing and Scientific prose. Each genre was further subdivided in two classes according to their expected target audience, taken as indicative of the accessibility level of the document. The journalistic genre class includes two different corpora: a newspaper corpus, *La Repubblica*, and an easy–to–read newspaper, *Due Parole* which was specifically written by linguists expert in text simplification using a controlled language for an audience of adults with a rudimentary literacy level or with mild intellectual disabilities (Piemontese, 1996). The Literature and Educational genre classes are partitioned into two subclasses, including texts respectively targeting children vs adults. The scientific prose genre class includes articles from Wikipedia as opposed to scientific articles. Among all these corpora, due to its peculiar nature *Due Parole* 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 1.

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). These corpora were automatically POS tagged by the Part–Of–Speech tagger described in (Dell'Orletta, 2009) and dependency–parsed by the DeSR parser (Attardi, 2006) using Support Vector Machine as learning algorithm.

For readability classification experiments we used READ–IT (Dell'Orletta et al., 2011b), the only available NLP–based readability assessment tool dealing with Italian texts. It uses lexical, morpho–syntactic and syntactic features, listed in Table 2, which are reliably identified from syntactically (i.e. dependency) parsed texts. It is a classifier based on Support Vector Machines that, given a set of features and a training corpus, creates a statistical model which is used in the assessment of readability of unseen documents.

Abbreviation name	Corpus	Coarse–grained genre	N.documents	N.words
Rep	*La Repubblica* (Marinelli et al., 2003), Italian newspaper	*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 1: Corpora.

3 Readability Classification Across Textual Genres

In order to explore whether and to what extent readability is related to the textual genre, we carried out two sets of experiments which are aimed at discerning within each of the four genre classes easy– vs difficult–to–read documents and which differ at the level of the used models: in the first set, we used a single statistical model for all four genres, whereas in the second set the classification task 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. Accuracy is a global score referring to the percentage of correctly classified documents whereas Precision and Recall have been computed with respect to the target classes: in particular, Precision is the ratio of the number of correctly classified

Feature category	Name
Raw Text	*Average number of word for sentence* *Average number of character for word*
Lexical	*Type/Token Ratio* *Lexical density*
Morpho–syntactic	*Part-Of-Speech unigrams* *Verbal mood*
Syntactic	*Distribution of dependency types* *Depth of the whole parse tree* *Average depth of embedded complement 'chains'* *Distribution of embedded complement 'chains' by depth* *Number of verbal roots* *Arity of verbal predicates* *Distribution of verbal predicates by arity* *Distribution of subordinate vs main clauses* *Relative ordering with respect to the main clause* the *Average depth of 'chains' of embedded subordinate clauses* the *Distribution of embedded subordinate clauses 'chains' by depth* *Length of dependency links feature*

Table 2: Feature set.

documents as belonging to one target class over the total number of documents classified as belonging to the same class; Recall has been computed as the ratio of the number of correctly classified documents of a given target class over the total number of documents belonging to the same class; F–measure is the weighted harmonic mean of Precision and Recall.

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 easiest to read 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 4). For the last and 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 3, the columns headed by *2Par/Rep Model*, *2Par/ScientArt Model* and *All Easy/All Difficult Model* show the results achieved for each testual 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, also used in (Dell'Orletta et al., 2011b), 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 (Dell'Orletta et al., 2011b) (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 sets.

Consider now the results of the second set of experiments carried out using a specific model for each of the four genres, reported in the Column headed *Genre–specific Models* in Table 3. As expected, the overall accuracies significantly increase with respect to the results obtained by the single 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 does not only include easy–to-read documents with respect to the *ScientArt* class: in fact, articles concerning a specific domain in Wikipedia can also include technical (i.e. difficult–to–read) documents. With this first set of experiments, we showed 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 for each genre should be built. This represents a difficult objective, especially in real–world applications. In what follows, a possible 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.

	2Par/Rep Model			2Par/ScientArt Model			All Easy/All Difficult Model			Genre–specific Models		
Genre	Prec	Rec	F–measure	Prec	Rec	F–measure	Prec	Rec	F–measure	Prec	Rec	F–measure
2Par	100	96.67	98.30	50.85	100	67.41	93.55	96.67	95.08	100	96.67	98.30
Rep	96.78	100	98.36	100	3.33	6.45	96.55	93.33	94.91	96.77	100	98.36
	Accuracy: 98.33			**Accuracy: 51.67**			**Accuracy: 95**			**Accuracy: 98.33**		
ChildLit	0	0	0	46.81	73.33	57.14	100	46.67	63.63	84.61	73.33	78.57
AdLit	50	100	66.67	38.46	16.67	23.25	65.22	100	78.95	76.47	86.67	81.25
	Accuracy: 50			**Accuracy: 45**			**Accuracy:73.33**			**Accuracy: 80**		
ChildEdu	90	31.03	46.15	49.15	100	65.91	56.67	58.62	57.63	78.79	89.65	83.87
AdEdu	59.18	96.67	73.42	0	0	0	58.62	56.67	57.63	88.46	76.67	82.14
	Accuracy: 64.41			**Accuracy: 49.15**			**Accuracy: 57.63**			**Accuracy: 83.05**		
Wiki	100	20	33.33	81.25	86.67	83.87	47.17	83.33	60.24	53.57	100	69.77
ScientArt	55.55	100	71.43	85.71	80	82.76	28.57	6.67	10.81	100	13.33	23.53
	Accuracy: 60			**Accuracy: 83.33**			**Accuracy: 45**			**Accuracy: 56.67**		
TOT Easy–to–access	97.78	36.97	53.66	54.31	89.91	67.72	66.40	71.43	68.82	74.30	89.91	81.37
TOT Difficult–to–access	61.34	99.17	75.80	71.43	25	37.04	69.34	64.17	66.67	87.37	69.17	77.21
	Accuracy: 68.20			**Accuracy: 57.32**			**Accuracy: 67.78**			**Accuracy: 79.51**		

Table 3: Classification–based readability assessment results.

4 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 (listed in Table 2). The readability score is computed as a linear combination between the distance of an analysed document (d) and two n–dimensional vectors representing 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 (Tanaka-Ishii et al., 2010) ranking method based on a *comparison* strategy.

As stated in Section 2, we assumed the vector representing the training set of *Due Parole* as the easiest–to–read pole, while the difficult–to-read extreme was selected computing the cosine distance of the vector representing each of the eight training sets (resulting from the genre/readability combination) from the *2Par* vector. The *ScientArt* vector turned out to be the most distant one and for this reason it was chosen as the difficult extreme. We report below the ordered list of the test set vectors ranked according to their readability scores:

$$\text{2Par} < \text{EduInf} < \text{LitInf} < \text{Rep} < \text{Wiki} < \text{LitAd} < \text{AdEdu} < \text{ScientArt}$$

Note that the relative order between the easy– and difficult–to–read subclasses for each genre

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.

Table 4 reports the ranking of all test documents based on the distance readability score. Each row represents a set of 30 documents. Interestingly, for each genre class the number of easy–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. Consider, for instance, the easy–to–read test sets: whereas for *2Par*, *ChildLit* and *ChildEdu* the distribution across the 30–document groups follows the expectations, *Wiki* documents are homogeneously distributed in all classes. Similar observations hold in the case of the *Rep* documents for what concerns the difficult–to–read class.

	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 4: Ranking–based readability assessment results.

The results of the ranking–based readability assessment method can be used as such but can also be exploited to create genre–specific training sets which, as demonstrated in Section 3, are needed to achieve reliable results in a classification–based readability assessment task. In order to test reliability and effectiveness of our ranking method for the automatic construction of training datasets, we focused on the *Scientific writing* genre for which we obtained the most unsatisfactory results. 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 5 reports the results of READ–IT with the new genre–specific model, using the automatically constructed *Wiki* training set: with respect to the previous genre–specific model, we obtained an improvement of 21.66% in Accuracy, thus demonstrating effectiveness and reliability of the proposed ranking method.

Genre	Prec	Rec	F–measure
Wiki	72.97	90	80.60
ScientArt	86.96	66.67	75.47
	Accuracy: 78.33		

Table 5: Classification results on *Scientific prose* using the automatically revised training set.

Conclusion

In this paper, we have shown that readability assessment is strongly influenced by textual genre and for this reason a genre–oriented notion of readability is needed. This represents an important requirement as far as educational applications are concerned. In particular, we demonstrated that with classification–based approaches to readability assessment reliable results can only be achieved with genre–specific models: this is far from being a workable solution, especially for less–resourced languages. We also proposed a new ranking method for readability assessment based on the notion of distance, which can be usefully exploited for automatically building genre–specific training corpora.

References

Aluisio, S., Specia, L., Gasperin, C., and Scarton, C. (2010). Readability assessment for text simplification. In *Proceedings of the 2NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9.

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

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

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

Dell'Orletta, F., Montemagni, S., and Venturi, G. (2011b). Read-it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the the Workshop on "Speech and Language Processing for Assistive Technologies" (SLPAT 2011)*, pages 73–83, Edinburgh, July 30.

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

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

Kate, R. J., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R. J., Roukos, S., and Welty, C. (2010). Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, pages 546–554.

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

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

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

Nenkova, A., Chae, J., Louis, A., , and 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. T., editor, *Empirical Methods in NLG*, pages 222–241, Berlin Heidelberg. LNAI 5790, Springer-Verlag.

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

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

Tanaka-Ishii, K., Tezuka, S., and Terada, H. (2010). Sorting texts by readability. In *Comput. Linguist.*, pages 203–227, 36, 2. MIT Press, Cambridge, MA, USA.

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