Linguistic Profiling of Texts Across Textual Genres and Readability Levels. An Exploratory Study on Italian Fictional Prose

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

Abstract

In this paper we present a case study focusing on the literature genre, in particular on Italian fictional prose, aimed at identifying the features characterizing this text type. Identified features were tested in two classification tasks, i.e. by genre and by readability, with promising results. Interestingly, the same multi–level set of linguistic features turned out to reliably capture variation within and across textual genres.

1 Introduction

Over the last ten years, Natural Language Processing (NLP) techniques combined with machine learning algorithms started being used to investigate the "form" of a text rather than its content. The range of tasks sharing this approach to the analysis of texts is wide, ranging e.g. from native language identification (see among the others Koppel et al. (2005) and Wong and Dras (2009)), author recognition and verification (see e.g. van Halteren (2004), authorship attribution (see Juola (2008) for a survey), genre identification (Mehler et al., 2011) to readability assessment (see Dell'Orletta et al. (2011a) for an updated survey). Besides obvious differences at the level of selected linguistic features and learning techniques, which are also motivated by the language varieties targeted by the different tasks, they share a common approach: they succeed in determining the language variety, the author, the text genre or the level of readability of a text by exploiting the distribution of features automatically extracted from texts. The issues typically dealt with in this type of studies can be summarised in two main research questions aimed at investigating 1) which linguistic features work best for a given task, and 2) which type of machine learning algorithms are best suited for a given task.

In this paper, we focus on the first issue, i.e. on the typology of linguistic features which could be reliably extracted from automatically analysed texts with particular attention to the potential impact of achieved results on two classification tasks. In particular, we identified the set of linguistic features characterizing classes of documents, based on their textual genre or the type of audience they target: to put it in van Halteren words (van Halteren, 2004), we carried out "linguistic profiling" of texts selected as representative of different genres and/or readability levels. Achieved theoretical results were tested in two text classification tasks, aimed at classifying texts by genre or readability level. This goal was pursued in a case study focusing on the literature genre, in particular on Italian fictional prose. First, we studied variation within and across genres, by carrying out a contrastive linguistic analysis a) of a corpus of literature texts with respect to corpora representative of other textual genres, and b) within the class of literary texts based on the expected target audience (adult vs children). Second, identified features were exploited as a proof of concept in two classification tasks, aimed at automatically discriminating literature texts from texts belonging to other genres, and literature texts targeting adults vs children. A qualifying feature of our approach to the problem consists in the fact that the set of linguistic features explored to capture variation within and across textual genres is wide and, thanks to the most recent developments of NLP technologies, covers different levels of linguistic description, including syntax. The selection of features was not driven by the specific task we had in mind: we show that the same set of features turned out to be appropriate for two different and quite unrelated tasks such as genre classification and readability assessment. According to the most recent literature on readability, the degree of readability appears to be, at least to some extent, connected

Proceedings of Recent Advances in Natural Language Processing, pages 189–197,
Hissar, Bulgaria, 7-13 September 2013.

to the textual genre of the document under evaluation (Kate, 2010; Štajner, 2012; Dell'Orletta et al., 2012): linguistic features correlated with readability are also genre dependent. In particular, the results achieved in this case study are in line with those obtained by (Sheehan, 2013) who 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, and that the accuracy of readability predictions can be improved by using genre-specific models (this is also claimed by (Dell'Orletta et al., 2012)).

2 Linguistic Features

As Biber and Conrad (2009) put it, linguistic varieties – which they qualify as "registers" from a functional perspective – differ "in their characteristic distributions of pervasive linguistic features, not the single occurrence of an individual feature". This is to say that by carrying out the linguistic analysis of a variety, e.g. a textual genre, we need to quantify the extent to which a given feature occurs. Differences lie at the level of the distribution of linguistic features, which can be common and pervasive in some varieties but comparatively rare in others: e.g. the relative distribution of nouns and pronouns differs greatly between texbooks and literature (the former have fewer pronouns and more repetitions of nouns, while fiction shows a greater use of pronouns). For the specific concerns of this study, we focused on a wide set of features ranging across different linguistic description levels which are typically used in studies focusing on the "form" of a text, e.g. on issues of genre, style, authorship or readability. This represents a peculiarity of our approach: we resort to general features qualifying the lexical and grammatical characteristics of a text, rather than ad hoc features, specifically selected for a given text type or task. This choice makes the selected features highly domain–independent and portable across different tasks (see Section 5).

The set of selected features is described below, organised into four main categories defined on the basis of the different levels of linguistic analysis automatically carried out (tokenization, lemmatization, morpho–syntactic tagging and dependency parsing): i.e. raw text features, lexical features as well as morpho-syntactic and syntactic features.

Raw Text Features

They include *Sentence Length*, calculated as the average number of words per sentence, and *Word Length*, calculated as the average number of characters per word.

Lexical Features

Basic Italian Vocabulary rate features: they refer to the internal composition of the vocabulary of the text. As a reference resource we took the *Basic Italian Vocabulary* by De Mauro (2000), including a list of 7000 words highly familiar to native speakers of Italian. In particular, we calculated two different features corresponding to: *i)* the percentage of all unique words (types) on this reference list (calculated on a per–lemma basis); *ii)* the internal distribution of the occurring basic Italian vocabulary words into the usage classification classes of 'fundamental words' (very frequent words), 'high usage words' (frequent words) and 'high availability words' (relatively lower frequency words referring to everyday life).
Type/Token Ratio: the Type/Token Ratio (TTR) is a measure of vocabulary variation which has shown to be helpful for measuring lexical variety within a text. Due to its sensitivity to sample size, TTR has been computed for text samples of equivalent length (the first 1000 tokens).

Morpho–syntactic Features

Distribution 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.
Lexical density: it refers to the ratio of content words (verbs, nouns, adjectives and adverbs) to the total number of lexical tokens in a text.
Mood, tense and person of verbs: this complex feature refers to the distribution of verbs according to their mood, tense and person. It is a central feature in a language like Italian, characterized by a rich verbal morphology.

Syntactic Features

Distribution of dependency types: this feature refers to the distribution of different types of syntactic dependencies (e.g. subject, direct object, modifier, etc.).
Parse tree depth features: tree depth is indicative of sentence complexity as stated by, among

others, Yngve (1960), Frazier (1985) and Gibson (1998). This set of features 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 *the distribution of embedded complement 'chains' by depth*.

Verbal predicates features: these features capture different aspects of the behaviour of verbal predicates and include a) the *number of verbal roots* with respect to number of all sentence roots occurring in a text, b) their arity calculated as the number of instantiated dependency links sharing the same verbal head (covering both arguments and modifiers), c) the *distribution of verbal predicates by arity* and d) the *percentage of verbal predicates with elliptical subject* (Italian is a pro–drop language). Concerning b), we believe that both a low and a high number of dependents can represent peculiar features of a given linguistic variety, corresponding to elliptical constructions in the former case and to a high number of modifiers (locative, temporal, manner, etc.) in the latter.

Subordination features: Features in this class include: a) the *distribution of subordinate vs main clauses*; b) 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 process); c) the *average depth of 'chains' of embedded subordinate clauses*; and d) the *the distribution of embedded subordinate clauses 'chains' by depth*.

Length of dependency links: Lin (1996) and Gibson (1998) showed that the syntactic complexity of sentences can be predicted with measures based on the length of dependency links. We measure the dependency length in terms of the words occurring between the head and the dependent.

3 Corpora and Pre–processing Tools

Four corpora representative of traditional textual genres, i.e. Literature, Journalism, Educational writing and Scientific prose, are considered. These corpora (detailed in Table 1) are internally subdivided into two different sets, according to the expected target audience. In particular, the journalistic corpus is articulated into a newspaper corpus, *La Repubblica*, and an easy–to–read newspaper corpus, *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 Educational corpus is partitioned into two subclasses, including texts targeting primary school vs high school. The scientific prose corpus includes articles from Wikipedia as opposed to scientific articles. For what concerns the Literature genre, we focused on one of the three major literary genres, namely fictional prose. In particular, the corpus of Italian literary texts explored here is subdivided into two different sub–corpora, constituted by adult and children literature respectively. The adult literature corpus is part of the Italian PAROLE Corpus (Marinelli et al., 2003) and includes 44 novels, either written by Italian writers or Italian translations of foreign novels (very few cases), published between 1974 and 1989. The children literature corpus is part of the wider corpus used for building a statistically–based children's lexicon (Marconi et al., 1994) and includes novels whose target are children of the primary school.

All corpora were automatically morpho-syntactically tagged by the POS tagger described in Dell'Orletta (2009) and dependency–parsed by the DeSR parser (Attardi, 2006) using Support Vector Machine as learning algorithm. DeSR, trained on the ISST–TANL treebank consisting of articles from newspapers and periodicals, achieves a performance of 83.38% and 87.71% in terms of LAS and UAS respectively when tested on texts of the same type (Attardi et al., 2009). However, since Gildea (2001) it is widely acknowledged that parsers have a drop of accuracy when tested against corpora differing from the typology of texts on which they were trained. Therefore, we can assume that the performance of DeSR is probably worse when parsing texts belonging to a different textual genre, such as literature or scientific writing. Despite this fact, we expect that useful information can be extracted from the linguistically annotated text, especially for what concerns the way lexical and grammatical patterns instantiating the features described in Section 2 recur across different text types.

Genre	Corpus	N.documents	N.words
Literature	*Children Literature* (Marconi et al., 1994)	101	19,370
	Adult Literature (Marinelli et al., 2003)	327	471,421
		Total: 428	**Total: 490,791**
Journalism	*La Repubblica* (Marinelli et al., 2003), Italian newspaper	321	232,908
	Due Parole, easy–to–read Italian newspaper (Piemontese, 1996)	322	73,314
		Total: 643	**Total: 306,222**
Educational	*Educational Materials* for Primary School (Dell'Orletta et al., 2011b)	127	48,036
	Educational Materials for High School (Dell'Orletta et al., 2011b)	70	48,103
		Total: 197	**Total: 96,139**
Scientific prose	*Wikipedia* articles from the Italian Portal "Ecology and Environment"	293	205,071
	Scientific articles on different topics (e.g. climate changes and linguistics)	84	471,969
		Total: 377	**Total: 677,040**

Table 1: Corpora.

4 Linguistic Profiling Results

4.1 Linguistic Profiling across Genres

In this section, we discuss a selection of linguistic profiling results corresponding to some of the features which turned out to strongly characterize the Literature genre with respect to the other textual genres taken into account. Starting from raw textual features, it can be noticed (see Table 2) that both average sentence length and average word length show much lower values if compared with the other corpora: this is in line with the Biber and Conrad (2009)'s claim that words and sentences in scientific writing as well as in other types of highly informative texts are much longer than fictional prose where short and simple words are typically used instead of long technical terms. Among the lexical features, the Literature genre appears to record the higher TTR value, meaning that this text type is characterized by a greater lexical variety. For what concerns morpho–syntactic features such as Part–of–Speech distribution, literary texts show a higher occurrence of pronouns and verbs, two features which are more common in conversation than in written language varieties (Biber and Conrad, 2009). On the other hand, quite a low frequency of occurrence of nouns can be observed, giving rise to a much lower noun/verb ratio. Following Voghera (2005) this can be explained in different ways: first, differently from informative texts fictional prose can have dialogical parts, which presumably present a distribution of nouns and verbs closer to that of spoken language; secondly, novels have long narrative parts in which the progression of the text leads to chains of verbal clauses, and this is crucial to determine a higher frequency of verbs. Other important features of fictional prose concern the use of subordinating constructions. This tendency comes out clearly from the different linguistic annotation layers: at the level of morpho–syntax we can observe a higher occurrence of subordinative conjunctions (as opposed to coordinative conjunctions) with respect to the other genres; at the dependency annotation level a higher percentage of subordinate clauses (as opposed to main clauses) is registered, which is also confirmed by the highest average depth of embedded subordinated constructions associated with the literature genre. This strong tendency towards the use of subordination is reminiscent of spoken language which commonly relies on dependent clauses embedded in higher level clauses: e.g. *that* complement clauses controlled by a verb and finite adverbial clauses (e.g. *because*– or *if*–clauses) which are actually much more common in conversation than in informative writing (Biber and Conrad, 2009). Other features which fictional prose shares with spoken language but make it differ from other genres are concerned with the use of ellipsis (see the lower percentage of verbal roots with explicit subject) and of verbal tense (see the lower occurrence of present tense verbs and the high frequency of past tense verbs).

4.2 Linguistic Profiling of Child vs Adult Literature Corpora

In spite of the fact that when compared with other textual genres the Literature corpus taken

Features	*Lit*	*Jour*	*ScientArt*	*Edu*
Average sentence length	17.99	22.90	27.19	28.15
Average word length	4.91	5.09	5.57	5.00
Type/token ratio (first 100,000 tokens)	0.71	0.63	0.66	0.69
Distribution of Parts–Of–Speech:				
– nouns	23.63	28.29	28.53	23.25
– verbs	15.20	13.30	10.67	13.87
– pronouns	6.32	3.05	3.12	5.42
Noun/verb ratio	1.55	2.13	2.67	1.68
Internal distribution of conjunctions:				
– subordinating	29.80	21.60	16.21	21.71
– coordinating	70.20	78.40	83.79	78.29
Distribution of verb tense:				
– simple present	36.26	55.63	54.33	40.67
– simple past	9.79	1.02	1.40	7.27
– imperfect	17.01	4.68	1.27	15.32
Average length of the longest dependency link	7.26	9.11	10.37	10.91
Average parse tree depth	4.57	5.91	6.74	6.57
Average depth of embedded complement 'chains'	1.17	1.30	1.38	1.22
Main vs subordinate clauses distribution:				
– main clauses	66.53	70.55	72.26	67.01
– subordinate clauses	33.23	29.30	27.47	32.23
Average depth of 'chains' of embedded subordinate clauses	1.14	1.09	0.96	1.09
Distribution of verbal roots with explicit subject	48.79	69.70	76.60	66.90

Table 2: An excerpt of linguistic profiling results.

as a whole has a peculiar linguistic profile which makes it significantly different from the other genres, the genre–internal analysis of children vs adult literary texts shows systematic differences. For illustrative purposes, the results of this genre–internal analysis have been compared with a corpus representative of another genre in order to show that in spite of the recorded differences the peculiarities of the literature genre are still clear and visible. We selected to this end Scientific prose, which turned out to be the most distant genre from Literature. Starting from the analysis of the lexical features, it can be noticed that the corpus of texts targeting children (henceforth, *ChildLit*) differs from the collection of texts addressing adults (henceforth, *AduLit*). As Table 3 shows, the *ChildLit* corpus contains a higher percentage of lemmas (types) belonging to the "Basic Italian Vocabulary" (BIV in the table) with respect to the *AduLit* corpus. This is in line with the outcomes of the studies on the discriminative power of vocabulary clues in a readability assessment task (see, among others, Petersen and Ostendorf (2009)): it witnesses the efforts of the authors of children books towards the use of a simple and comprehensible vocabulary. In spite of these differences, a more extended use of basic vocabulary is observed in the literature as a whole with respect to the *ScientArt* corpus characterized by a much lower percentage of BIV words. At the syntactic level, the *ChidLit* and *AduLit* corpora are characterized by different complexity levels. *AduLit* contains *i)* sentences longer than those occurring in the books for children, *ii)* the highest percentage of long dependency links as well as the deepest syntactic trees, and *iii)* the highest percentage of complex nominal constructions with deep sequences of embedded complements. Conversely, for what concerns *iii)*, *ChidLit* is characterized by: a higher percentage of short sequences, i.e. with depth=1 (83.18%) with respect to *AduLit* (77.16%); a lower percentage of sequences of embedded complement chains with depth=$\geq$ 3, covering only 1.73% of all 'chains' as opposed to 2.64% in *AduLit*. Despite these genre–internal differences, the lower syntactic complexity level of the literature with

Features	*ChildLit*	*AduLit*	*ScientArt*
Average sentence length	16.96	18.25	27.19
% of lemmas (types) in BIV	73.95	69.57	58.54
% of lemmas (types) NOT in BIV	26.05	30.43	41.46
Distribution of Parts–Of–Speech:			
– nouns	21.96	24.08	28.53
– verbs	15.83	14.96	10.67
– pronouns	6.88	6.13	3.12
Average length of the longest dependency link	6.63	7.43	10.37
Average parse tree depth	4.51	4.57	6.74
Distribution of 'chains' by depth:			
– 1 embedded complement	83.18	77.16	69.77
– 2 embedded complements	14.11	15.61	22.66
$\geq$ 3 embedded complements	1.73	2.64	7.05
Main vs subordinate clauses distribution:			
– main clauses	68.32	65.77	72.26
– subordinate clauses	30.69	33.92	22.47
Distribution of post–verbal subordinate clauses	88.54	81.16	78.55
Distribution of verbal roots with explicit subject	52.33	47.54	76.60

Table 3: An excerpt of features discriminating adult from children literature corpora.

respect to the scientific prose genre is still visible: *ScientArt* contains longer dependency links, higher syntactic trees and deeper sequences of embedded complements. As seen in Section 4.1, a further qualifying feature of the literary genre is the recurrent use of subordination, which occurs much less frequently in the *ScientArt* corpus. In *ChildLit* subordinate clauses represent the 30.69% of the total amount of clauses occurring in the corpus and they mostly follow the main clause, i.e. 88.54% of the subordinate clauses occur in post–verbal position, while subordinated clauses represent 33.92% of the clauses in the *AduLit* corpus and occur less frequently (81.16%) in post–verbal position. This can be taken as a further proof of the higher syntactic complexity of the *AduLit* corpus. According to the literature, the use of parataxis is preferable to a hypotactic structure since a coordinated construction is in principle more easy–to–read and comprehensible than a subordinate one (Beaman, 1984; Piemontese, 1996). The higher number of post–verbal subordinates in *ChildLit* is in line with Miller and Weinert (1998) claim that subordinate clauses occurring in post–verbal rather than in pre–verbal position are easier to process. Among the features concerning verbal predicates, the distribution of verbal roots with explicit subject, 52.33% in *ChildLit* and 47.54% in *AduLit*, can be indicative of a greater occurrence of elliptical constructions in the adult literature: this represents a peculiarity of literary texts which show a stronger tendency towards the ellipsis of grammatical elements.

5 Two Classification Tasks

5.1 Automatic Textual Genre Assessment

In order to explore whether and to what extent the features illustrated in Section 2 can be successfully exploited in an automatic genre classification task, the four corpora were randomly split into training and test sets. For each corpus, the test sets consist of 30 documents while the training sets include the following numbers of documents: 368 (Literature), 583 (Journalistic), 137 (Educational writing), 317 (Scientific prose). We built a classifier based on Support Vector Machines using LIBSVM (Chang and Lin, 2001) and we used two different models of features: a **Lexical Model**, using a combination of *raw text* and *lexical* features and a **Syntax Model**, combining all feature types. Achieved results have been evaluated in terms of *i)* overall Accuracy of the system and *ii)* Precision, Recall and F–measure. Table 4 reports the results achieved with the two models. The *Syntax Model* shows a significant improvement at the level of the accuracy score with respect to the

	Lexical model (**Accuracy: 62.18**)			Syntax model (**Accuracy: 76.47**)		
Genre	Prec	Rec	F–measure	Prec	Rec	F–measure
Journalism	44.64	83.33	58.14	61.63	88.33	72.60
Literature	77.59	76.27	76.92	85.71	91.52	88.52
Educational	80	6.77	12.5	92.59	42.37	58.14
Scientific prose	77.78	81.67	79.67	80.64	83.33	81.97

Table 4: Genre classification results.

Lexical Model, demonstrating that when the aim is capturing the "form" of a text a crucial role is played by morpho–syntactic and syntactic features, which also play a significant role in the linguistic profiling of texts. It can be noted that, using the *Syntax Model*, the classification of the documents in the class *Literature* achieves a higher F–measure (88.52%) with respect to the *Educational* class which shows the lowest F–measure value (58.14%). We can hypothesize that, as reported in Table 2, the *Literature* genre is strongly characterized with respect to the other textual genres considered here. The fictional prose documents show a strong tendency towards, for example, short dependency links, shallow syntactic trees as well as towards a low percentage of verbal roots with explicit subjects. On the contrary, the results achieved with respect to the *Educational* texts can follow from the internal composition of this corpus gathing a heterogeneous collection of documents (such as textbooks, anthologies, exercises, etc.): this fact may have negatively affected the classification accuracy of the *Educational* texts.

5.2 Automatic Readability Assessment

Starting from the assumption that the expected target audiences of *ChildLit* and *AduLit* texts can be taken as indicative of their accessibility level, we modeled the task of automatically discriminating between children and adult literature as a genre–specific automatic readability assessment task. For this purpose, we used READ–IT (Dell'Orletta et al., 2011a), the only available NLP–based readability assessment tool for Italian. READ–IT exploits the wide typology of lexical, morpho–syntactic and syntactic features illustrated in Section 2. As in the previous case, the classifier is based on SVM that, given a set of features and a training corpus, creates a statistical model which is used for assessing the readability of unseen documents. In this experiment, the *ChildLit* and *AduLit* corpora were split into training and test sets. For each of them, the test sets consist of 30 documents, whereas the training sets include respectively 71 and 297 documents. Achieved results are evaluated in terms of overall Accuracy, Precision, Recall and F–measure. As shown in Table 5, READ–IT performs better at the level of F–measure in the classification of *AduLit* rather than of *ChildLit* texts. As discussed in (Dell'Orletta et al., 2012), this may follow from the small amount of training data available for the children literature class. However, interestingly enough, even if the *AduLit* and *ChildLit* training sets have quite different sizes, the variation internal to the genre was successfully captured by the classifier which achieves an overall Accuracy of 80%. Achieved results show that the set of selected features is also able to reliably capture genre–internal variation.

	Prec	Rec	F–measure
ChildLit	84.61	73.33	78.57
AdLit	76.47	86.67	81.25
	Accuracy: 80		

Table 5: Readability assessment results.

6 Conclusion

In this paper we reported the results of a case study focusing on the literature genre and aimed at carrying out "linguistic profiling" of literary texts as opposed to other textual genres such as Journalism, Educational writing and Scientific prose. Achieved theoretical results concerning the linguistic characterization of the genre represented by Italian fictional prose are nicely complemented by applicative results showing that the features identified can be reliably put at work in two text classification tasks, i.e. the automatic assessment of textual genre and readability level. Interestingly, the same multi–level set of linguistic features was used to capture variation within and across textual genres, without any ad hoc selection of features. Current developments include feature selection and ranking for both genre classification and readability assessment tasks.

References

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

Giuseppe Attardi, Felice Dell'Orletta, Maria Simi, Joseph Turian. 2009. Accurate Dependency Parsing with a Stacked Multilayer Perceptron. In: Proceedings of Evalita'09.

Douglas Biber. 1993. Using Register–diversified Corpora for General Language Studies. *Computational Linguistics Journal*, 19(2): 219–241.

Douglas Biber and Susan Conrad. 2009. Genre, Register, Style. Cambridge: CUP.

Karen Beaman. 1984. Coordination and Subordination Revisited: Syntactic Complexity in Spoken and Written Narrative Discourse. In Tannen D. and Freedle R. (eds), *Coherence in Spoken and Written Discourse*, 45–80.

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

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

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

Felice Dell'Orletta, Simonetta Montemagni, Eva Maria Vecchi and Giulia Venturi. 2011b. 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*, McGraw–Hill, 319–336.

Felice Dell'Orletta, Simonetta Montemagni and Giulia Venturi. 2011a. READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification. In *Proceedings of the Workshop on "Speech and Language Processing for Assistive Technologies" (SLPAT 2011)*, Edinburgh, July 30, 73–83.

Felice Dell'Orletta, Simonetta Montemagni and Giulia Venturi. 2012. Genre–oriented Readability Assessment: a Case Study. In *Proceedings of the Workshop on Speech and Language Processing Tools in Education (SLP-TED)*, 91–98.

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

David Elson, Anna Kazantseva, Rada Mihalcea and Stan Szpakowicz (eds.). 2012. Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature. Montréal, Canada, June 2012, Association for Computational Linguistics, http://www.aclweb.org/anthology/W12-25.

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

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

Daniel Gildea. 2001. Corpus variation and parser performance. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP 2001)*, 167–202.

Patrick Juola. 2008. *Authorship Attribution*. Now Publishers Inc.

Rohit J. Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J. Mooney, Salim Roukos and Chris Welty. 2010. Learning to Predict Readability using Diverse Linguistic Features. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, 546–554.

Moshe Koppel, Jonathan Schler and Kfir Zigdon. 2005. 2005. Automatically determining an anonymous author's native language. In *Intelligence and Security Informatics*, vol. 3495, LNCS, Springer–Verlag, 209–217.

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

Lucia Marconi, Michela Ott, Elia Pesenti, Daniela Ratti and Mauro Tavella. 1994. *Lessico Elementare*. Zanichelli, Bologna.

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

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

Alexander Mehler, Serge Sharoff and Marina Santini (Eds.). 2011. *Genres on the Web. Computational Models and Empirical Studies*. Springer Series: Text, Speech and Language Technology.

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

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

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

Sze–Meng Jojo Wong and Mark Dras. 2009. Contrastive Analysis and Native Language Identification. In *Proceedings of the Australasian Language Technology Association Workshop*.

Kathleen M. Sheehan, Michael Flor and Diane Napolitano. 2013. A Two-Stage Approach for Generating Unbiased Estimates of Text Complexity. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, 49-58.

Sanja Štajner, Richard Evans, Constantin Orasan and Ruslan Mitkov. 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)*.

Hans van Halteren. 2004. Linguistic profiling for author recognition and verification. In *Proceedings of the Association for Computational Linguistics (ACL04)*, 200–207.

Miriam Voghera. 2005. Nouns and Verbs in Speaking and Writing. In E. Burr (eds.), *Tradizione e innovazione. Il parlato: teoria - corpora- linguistica dei corpora*, Firenze, Cesati, 2005, 485–498.

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