NLP–Based Readability Assessment of Health–Related Texts: a Case Study on Italian Informed Consent Forms

Giulia Venturi[◇], **Tommaso Bellandi**[•], **Felice Dell'Orletta**[◇], **Simonetta Montemagni**[◇]
[◇]Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC–CNR)
ItaliaNLP Lab - *www.italianlp.it*
{name.surname}@ilc.cnr.it
[•]Laboratorio per le attività di studio e ricerca applicata, Centro Gestione Rischio Clinico e Sicurezza dei Pazienti, Patient Safety Research Lab
bellandit@aou-careggi.toscana.it

Abstract

The paper illustrates the results of a case study aimed at investigating and enhancing the accessibility of Italian health–related documents by relying on advanced NLP techniques, with particular attention to informed consent forms. Results achieved show that the features automatically extracted from the linguistically annotated text and ranging across different levels of linguistic description have a high discriminative power in order to guarantee a reliable readability assessment.

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 of health–related information which should be accessible to all members of the society, including people who have reading difficulties e.g. as a result of a low education level, or of language–based learning disabilities, or because the language of the text is not their native language (WHO, 2015). It is a widely acknowledged fact that poor communication between physician and patients predisposes to medical malpractice cases (Kohn et al., 2000). Patient safety is a global challenge since the evidence on the burden of adverse events emerged in the past 15 years. An estimate of 43 millions adverse events occur in one year globally, with more than 50% of preventable events (Jha et al., 2013). In Italy, the incidence is of 5.2% on in–hospital admissions (Tartaglia et al., 2012) and the direct cost related to the prolongation of the stay are up to 3bln Euros in one year, roughly 3% of the funds of the National Healthcare Service (Albolino et al., 2013). The indirect costs related to claims is also high, amounting up to 1bln Euro in one year.

For all these reasons, the medical community has always shown strong interest in the improvement of health–related information in terms of document quality and understandability. Studies carried out so far mainly focused on traditional readability assessment methods, such as e.g. the Flesch–Kincaid measure (Kincaid, 1975) for the English language or the GulpEase index for Italian (Lucisano and Piemontese, 1988). According to them, the readability of medical texts is assessed by relying on basic text features such as sentence and word length, the only ones which could be automatically extracted from texts when these measures were originally conceived.

Recently, concerns have been raised about the effectiveness of traditional readability indices in capturing linguistic factors related to text complexity (Gemoets et al., 2004; Clerehan et al., 2005). This follows from the fact that now it is possible to carry out readability assessment against linguistically annotated texts, i.e. enriched with detailed and multi–level linguistic information generated by Natural Language Processing (NLP) components. Providing complex scientific information in a way that is comprehensible to a lay person is thus a challenge that nowadays can be addressed by 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 (Dell'Orletta et al., 2014a). So far, very few attempts have been devoted to the use of advanced NLP techniques to assess the readability of health–related texts; to our knowledge, none of them deals with Italian.

In this paper, we report the first results of a case study aimed at assessing the readability of a corpus of Italian informed consent forms on the basis of NLP–enabled surface, lexical and syntactic features. Among health–related texts, we focused on informed consent forms since ineffective doctor–

Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (Louhi), pages 131–141,
Lisbon, Portugal, 17 September 2015.

patient communication is often due to a weak or lacking informed consent (Korenman et al., 2015).

The case study was carried out in the framework of the collaboration between the Institute of Computational Linguistics of the Italian National Research Council (ILC–CNR) and the Centre for Clinical Risk Management and Patient Safety (GRC) of the Tuscany region whose final goal is the development of advanced technologies to support the improvement of doctor–patient communication. In particular, it originates from the fact that in 2010 GRC was appointed to manage a communication and compensation program on adverse events, in order to improve the effectiveness and efficiency of claims management. Thanks to this programme, after 5 years the efficiency has strongly improved, with an estimate saving of 50millions Euro per year and a reduction of 5 months to close the claim. Yet, the number of claims is still stable and a recurrence of cases related to ineffective doctor–patient communication, often related to a weak or lacking informed consent, continues to be observed. The collaboration between ILC–CNR and GRC is aimed at creating the prerequisites for improvig the effectiveness of doctor–patient communication. This goal is pursued by designing and developing a writing tool for clinical practitioners which includes advanced functionalities for the evaluation of the quality of written documents and for supporting their simplification (whenever needed): the paper reports the results of preliminary investigations aimed at evaluating the readability of a wide corpus of documents presented to patients for informed consent, covering a wide range of clinical specialties and released by different healthcare trusts.

2 Background

It is a widely acknowledged fact that NLP techniques have an impact on the design of readability measures enabling to capture complex linguistic features with a significant gain in performance (François and Miltsakaki, 2012). However, differently from other application scenarios, little effort has been devoted so far in the biomedical domain to fully exploit NLP potentialities to evaluate the readability of health–related texts and to support clinical practitioners in the simplification, whenever needed, of the documents they produce. NLP–based readability assessment approaches reported so far for the biomedical domain differ with respect to: whether readability assessment is carried out as a classification task or in terms of ranking; the typology of features taken into account; the application within which readability assessment is carried out; and, last but not least, the language dealt with.

2.1 Methods and Features

Classification–based methods carry out readability assessment by assigning a given document to predefined readability classes: this is the case, for instance, of Kauchak et al. (2014) who built a machine learning classifier for predicting the difficulty of medical texts trained on a data set of aligned sentence pairs collected from English Wikipedia and Simple English Wikipedia. Interestingly, in the biomedical literature on readability assessment readability classes are typically restricted to two, i.e. easy vs. difficult. However, the main drawback of classification models is that they require training data, which may not exist, especially for a specific domain. An alternative to this method is represented by ranking–based approaches, positioning the document being analysed within a readability ranking scale: this approach is better suited for dealing with less resourced languages or to meet the needs of specific domains. In the biomedical domain, this method is adopted, among others, by: Kim et al. (2007), who developed a domain–specific approach to readability assessment calculating a distance score based on whether and to what extent text features of a test document differ from those of an easy sample (consisting in a collection of various web health information resources); or Zeng-Treitler et al. (2012) who, with the aim of improving the rank–based approach by Kim et al. (2007), used a wider set of lexical features also taking into account frequency information.

For what concerns the typology of features, NLP–based approaches proposed so far mainly focus on a combination of grammatical features, typically represented by the distribution of Parts–Of–Speech or of noun phrases, and lexical features, such as the distribution of domain terms with respect to domain–specific vocabularies, e.g. the Unified Medical Language System (UMLS) vocabulary. This is the case, e.g., of Proulx et al. (2013) who, by combining grammatical and vocabulary features, developed a tool specifically addressing the needs of clinicians and health ed-

ucators for both readability assessment and enhancement. Since vocabulary plays a key role in health text readability, the most important extensions taken into account are concerned with lexical features. Starting from the assumption that more frequent terms are also easier to understand, Zeng-Treitler et al. (2012) included among the lexical features the distribution of terms with respect to two general–purpose resources, i.e. the Penn Treebank (Marcus et al., 1999) and the Google's Web 1T 5–gram Version 1 with n–gram frequency counts (Brants and Franz, 2006). For what concerns grammatical information, readability assessment in the biomedical domain does not go beyond to the distribution of Parts–Of–Speech and/or noun phrases: i.e. to our knowledge none of the domain–specific methods proposed so far makes use of syntactic features that can be extracted from the output of a syntactic parser.

2.2 Applications and Languages

Readability assessment is tackled from various perspectives with different applications in mind, giving rise to different tasks ranging from discerning easy vs. difficult electronic health records (Zeng-Treitler et al., 2007b), consumer health web sites, patient blogs and patient educational material (Leroy et al., 2006), to the simplification of medical texts carried out by devising metrics that can help making health–related documents more comprehensible to consumers. Due to the central role of lexical features in determining the readability of health–related texts, lexical simplification turned out to be the most explored level of text simplification. Different methods were devised to make health documents more comprehensible to consumers by reducing vocabulary difficulty. Even if with some differences, all approaches rely on the identification of difficult words and their replacement with easier synonym words. For this purpose, both domain–specific (e.g. Unified Medical Language System (UMLS), open–access collaborative (OAC), consumer health vocabulary (CHV)) and general–purpose (WordNet synonyms and hyperonyms, Wiktionary definitions, frequency counts of words in Google Web Corpus) resources were used. This is the case of Zeng-Treitler et al. (2007a) who built a prototype text translator to simplify narrative reports in electronic health reports, and of Leroy et al. (2012) who developed a semi–automatic algorithm tested on patient materials available on–line whose original and simplified version was presented for evaluation to a medical librarian (to measure the *perceived* difficulty) and to laymen (to measure the *actual* difficulty). Kandula et al. (2010) defined a text simplification method relying on both semantic and syntactic features: following Siddharthan (2006)'s approach, their algorithm is articulated into three steps, i.e. sentences longer than 10 words are first splitted, then Part–Of–Speech patterns are identified, and transformational rules are applied to generate shorter sentences.

Readability metrics developed so far typically deal with English, with few attempts tackling other languages. The most prominent exception is represented by Swedish, for which a quantitative corpus analysis of a collection of radiology reports was carried out as a preliminary step towards the development of a Swedish text simplification tool (Kvist and Velupillai, 2013). Similarly to English, simplification algorithms for Swedish health–related documents were devised by relying on synonym replacement methods (Abrahamsson et al., 2014), or on automatic detection of out–of–dictionary words and abbreviations, or on compound splitting and spelling correction (Grigonyte et al., 2014). Initiatives carried out so far for what concerns Italian are based on traditional readability formulas. This is the case of the ETHIC (*Evaluation Tool of Health Information for Consumers*) project (Cocchi et al., 2014), aimed at developing an effective tool for biomedical librarians and health information professionals to assess the quality of produced documents and to support them in preparing texts of increasing quality, suitable and comprehensible for patients and consumers in general. The tool carries out text readability and lexical understandability evaluation by resorting to the GulpEase readability formula (Lucisano and Piemontese, 1988) and the Basic Italian Vocabulary (De Mauro, 2000). Another relevant case study dealing with different languages also including Italian is reported in Terranova et al. (2012), whose aim was to assess and improve the quality and readability of informed consent forms used in cardiology. Although readability assessment was carried out with traditional readability formulas to guarantee comparability of results across languages, the main novelty of this study is that the simplification of Italian consent forms was guided by a preliminary version of READ–

IT (Dell'Orletta et al., 2011), the first NLP–based readability assessment tool for Italian.

3 The Approach

Our approach to the assessment of readability of Italian health–related texts combines NLP–enabled feature extraction and state–of–the–art machine learning algorithms. In this case study, we chose to exploit a general–purpose readability assessment tool, represented by READ–IT (Dell'Orletta et al., 2011)[1], the first NLP–based readability assessment tool for Italian which combines traditional raw text features with lexical, morpho–syntactic and syntactic information (see Section 3.2). In READ–IT, analysis of readability is modelled as a classification task. In particular, readability classification is binary, i.e. it is based on a training set consisting of two corpora representative of difficult– vs. easy–to–read texts. The easy–to–read training set is represented by *Due Parole* ("2Par"), a newspaper 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 at the lexicon and sentence structure levels (Piemontese, 1996). For the selection of the difficult–to–read training set we opted for texts belonging to the same class, i.e. newspapers. In particular, we used the daily newspaper *La Repubblica* ("Rep"): even if widely read by many people in Italy, the national statistics on literacy skills report that 71% of the Italian people can hardly comprehend texts of medium difficulty such as the *Rep* articles.

Two other qualifying features of the READ–IT approach to readability assessment are worth reporting here, namely: *i)* readability is assessed by considering a wide range of linguistic characteristics automatically extracted from linguistically annotated texts, and *ii)* readability analysis is carried out at both document and sentence levels. As reported in section 2, readability assessment in the biomedical domain typically relies on linguistic features extracted from automatically PoS–tagged texts: instead, our approach also includes features extracted from syntactically (i.e. dependency) parsed texts, thus making it possible to monitor a wider variety of factors affecting the readability of a text. The set of features can be parameterized creating the prerequisites for specializing the readability assessment measure with respect to different target audiences, specific domains of knowledge or with respect the type of textual object, i.e. the document or individual sentences. Assessing readability at both document and sentence levels allows highlighting specific text portions which require reformulation with respect to the used vocabulary or to the grammatical structure (Dell'Orletta et al., 2014c). In fact, similarly to other application scenarios, also in the biomedical domain evaluating the readability of individual sentences represents an essential prerequisite for text simplification, to be carried out at both lexical and syntactic levels (Kandula et al., 2010). Despite that sentence readability assessment is a qualifying feature of READ–IT, in what follows we will focus on document readability only.

For the experiments reported in the paper, we used general purpose readability models trained on newspaper corpora. This was an unavoidable choice, due to the lack of domain–specific resources annotated with grade levels to be used as training data. Although this makes achieved results still preliminary, it was a way to test effectiveness and reliability of the method on health–related texts. For what concerns the evaluation of achieved readability assessment results, the target readers of *2Par* (i.e. the READ–IT easy–to–read pole) were taken as coinciding with the target reader of health–related texts: the underlying assumption is that the informed consents classified as difficult–to read for the *2Par* low literacy readers are really complex and need to be simplified. Obviously, when a version of READ–IT specialized for the biomedical domain will be released, a qualitative evaluation of results will be needed. Previous studies resorted to the Cloze test to validate the reliability of their results or integrated editing capabilities into the developed tools in order to receive feedback from end users. The work carried out by Kandula and Zeng-Treitler (2008) represents an exception. They assembled a panel of experts to evaluate the readability of 324 different typology of English health documents: the rated collection was meant to be used as a gold standard to evaluate readability metrics.

3.1 The corpus

For this case study, we collected a corpus of 583 documents, for a total of 607,677 word tokens,

[1] http://www.italianlp.it/demo/read-it/

Features	***2Par***	***2IC***	***Rep***
Average sentence length	19.20	16.06	26.54
Average word length	4.98	6.75	5.18
% of lemmas (types) in BIV	74.58	57.24	67.09
% of lemmas (types) NOT in BIV	25.42	42.76	32.91
Type/token ratio (first 100 tokens)	0.55	0.72	0.72
Distribution of Parts–Of–Speech:			
– nouns	29.30%	28.51%	27.19%
– verbs	13.66%	11.83%	12.89%
– adjectives	5.92%	9.26%	6.40%
– prepositions	15.28%	16.19%	16.41%
Noun/verb ratio	2.14	2.41	2.11
Average length of the longest dependency link	7.91	6.43	10.28
Average parse tree depth	5.29	4.86	6.51
Average depth of embedded complement 'chains'	1.24	1.31	1.34
Distribution of 'chains' by depth:			
– 1 embedded complement	79.40%	74.25%	72.32%
– 2 embedded complements	17.02%	21%	21.42%
$\geq$ 3 embedded complements	3.27%	4.73%	5.87%
Main vs subordinate clauses distribution:			
– main clauses			
– subordinate clauses	26.14%	25.30%	32.36%
Average clause length	9.81	11.29	10.12
Distribution of verbal roots with explicit subject	74.69%	57%	64.30%

Table 1: Selection of linguistic features strongly characterizing the *2IC* corpus.

constituted by the procedures and the documents for informed consents currently used in all the 16 healthcare trust of the Regional Healthcare Service (RHS) of Tuscany, namely 4 academic hospitals and 12 local healthcare authorities. The documents were partitioned into different groups, classified according to the clinical specialty and the document type (procedure or user guide). Henceforth, we will refer to this corpus as the "Italian Informed Consent Corpus" (2IC).

Table 1 reports a selection of linguistic features which turned out to strongly characterize the *2IC* corpus with respect to the journalistic *2Par* and *Rep* corpora. This analysis is meant to compare domain–specific (i.e. biomedical) and general purpose corpora with the final aim of detecting the main linguistic features characterizing the language used in informed consent forms. The features were extracted from the corpus automatically tagged by the part–of–speech tagger described in Dell'Orletta (2009) and dependency–parsed by the DeSR parser (Attardi, 2006).

Starting from raw textual features, it can be noticed that the *2IC* corpus is characterized by shorter sentences (calculated as the average number of words per sentence) and longer words (calculated as the average number of characters per word) if compared with the *2Par* and *Rep* corpora. Starting from the assumption underlying traditional readability formulas assuming that longer sentences are more grammatically complex than shorter ones and that longer words are less comprehensible than shorter ones, this result witnesses the efforts of the authors of informed consents towards the use of an unavoidably complex vocabulary used, however, in simpler syntactic constructions. Interestingly enough, this is confirmed by the values of lexical features. Among them, it is worth noting that with respect to both *2Par* and *Rep* informed consents contain quite a lower percentage of lemmas (types) belonging to the "Basic Italian Vocabulary" (De Mauro, 2000), marked as BIV in Table 1 and corresponding to a list of 7000 words highly familiar to native speakers of Italian. This is in line with the outcomes of the studies on the discriminative power of vocabulary clues in readability assessment (see, among others, Petersen and Ostendorf (2009)). Obviously, this also reveals the massive use of health–related words specific to this domain of knowledge and here still considered as out–of–vocabulary lemmas. In addition, *2IC* texts show a higher Type–Token Ratio (TTR) value (which is computed for the first 100 tokens of each document), meaning that this text type is much richer lexically, with values which are closer to what observed with respect to *Rep*, here considered as representative of the class of difficult–to–read texts.

Consider now the distribution of Parts–Of–Speech across the *2Par*, *Rep* and *2IC* corpora. In-

formed consents are characterized by a high percentage of adjectives, prepositions and nouns, and by a low percentage of verbs: this gives rise to a much higher noun/verb ratio. According to Biber (1993), such different distributions represent significant dimensions of variation across textual genres. In particular, the higher noun/verb ratio reveals that informed consent forms are more informative than newspaper articles (Biber and Conrad, 2009), while the higher occurrence of nouns and prepositions is strongly connected with their presence within embedded complement 'chains' governed by a nominal head and including either prepositional complements or nominal and adjectival modifiers. Similarly to *Rep* articles, health–related documents contain a high percentage of complex nominal constructions (*Average depth of embedded complement 'chains'* in Table 1) with deep sequences of embedded complements. This is also reflected at the level of the probability distribution of embedded complement 'chains' by depth: if on the one hand we observe a lower percentage of short sequences (i.e. with depth=1) with respect to *2Par* here taken as representative of easy–to–read texts, on the other hand – similarly to *Rep* – a higher percentage of longer sequences (i.e. with depth=2 and $=\geq 3$) is recorded.

Interestingly, however, besides complexity features such as "heavy" nominal constructions – possibly due to multi–word terminology – in informed consents low values are recorded for syntactic features typically associated with *structural* complexity, such as: parse tree depth, calculated in terms of the longest path from the root of the dependency tree to some leaf; length of dependency links, measured in terms of the words occurring between the syntactic head and the dependent; or the distribution of main vs. subordinate clauses. From this, we can conclude that language complexity in informed consent forms mainly lies at the level of *local* features of the parse tree. Other peculiar syntactic features of informed consents with respect to *2Par* and *Rep* are represented by longer clauses (*Average clause length* in Table 1, calculated as the average number of tokens per clause), and a lower percentage of verbal roots with explicit subject (calculated with respect to the total amount of verbal roots). For what concerns the latter, even if Italian is a pro–drop language, sentences characterized by elliptical constructions (e.g. verbal roots with explicit subjects) make a text more difficult–to–read and need to be simplified, as suggested in Barlacchi and Tonelli (2013).

3.2 Readability Assessment

For readability classification experiments we used READ–IT, the first NLP–based readability assessment tool devised for Italian. It 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 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 or sentences. The assigned readability level ranges between 0 (easy–to–read) and 100 (difficult–to–read) referring to the percentage probability for the unseen documents or sentences to belong to the class of difficult–to–read documents. The score assigned by READ–IT can thus be seen as a score of text difficulty.

As fully described by Dell'Orletta et al. (2011), the tool is trained on *2Par* (taken as representative of the class easy–to–read texts) and on *Rep* (representing the class of difficult–to–read texts) articles and it exploits the wide typology of raw text, lexical, morpho–syntactic and syntactic features summarized in Table 2. This proposed four–fold partition of features 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.

The set of features used to build the statistical model can be parameterized through a configuration file. This creates the prerequisites for customising the readability assessment measure with respect to the target audience or to the sublanguage of a specific domain. According to the different types of features considered, READ–IT assigns different readability scores using the following four feature models:

1. **Base Model**, relying on *raw text* features only;

Feature category	Name
Raw Text	Average number of words per sentence Average number of characters per word
Lexical	Type/Token Ratio Lexical density *Basic Italian Vocabulary (BIV)* (De Mauro, 2000) rate
Morpho–syntactic	Part-Of-Speech unigrams Mood, tense and person of verbs
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 Distribution of embedded subordinate clauses 'chains' by depth Length of dependency links feature

Table 2: Linguistic features used for readability assessment purposes.

2. **Lexical Model**, relying on a combination of *raw text* and *lexical* features;

3. **Syntax Model**, relying on *morpho–syntactic* and *syntactic* features;

4. **Global Model**, combining all feature types, namely *raw text*, *lexical*, *morpho–syntactic* and *syntactic* features.

4 Results and Discussion

In this section, we discuss the outcome of the readability assessment experiments carried out on the *2IC* corpus described in Section 3.1. In order to identify the contribution of the different types of features in the assessment of the readability of informed consents, we focus on the results obtained by the base, lexical and syntactic READ–IT models (see, respectively, columns *Base*, *Lexical* and *Syntax* in Table 3). In what follows, we will focus on the results of the readability experiments carried out at the document level while an in depth investigation of the linguistic aspects affecting sentence readability is part of an on–going study.

Table 3 reports the results obtained with respect to the whole corpus, for all the 29 medical specialties; a score for each of the 4 considered macro–specialties (namely, *Surgery, Internal Medicine, Prevention and Medical Services*) is also computed, as the average of the scores recorded for each specialty. It can be noted that the whole corpus is characterized by a low readability level, even if with significant differences among the different readability models and across macro–specialties. Interestingly, the results obtained by the *Base* model show how raw text features such as sentence and word length are not really effective to capture the difficulty of these texts as well as the differences among them. This model can be seen as an approximation of the GulpEase index (Lucisano and Piemontese, 1988), i.e. the most used traditional readability measure for Italian which is based on the same raw text features (i.e. sentence and word length). This naturally follows from the results illustrated in Section 3.1, investigating the linguistic features characterizing *2IC* with respect to general purpose corpora: as Table 1 shows, *2IC* contains quite short sentences, a raw text feature typical of easy–to–read texts.

By comparing the scores obtained for the macro–specialties, it is worth noting that the score obtained with the *Base* model for the *Prevention* area is misleading: i.e, the prevention forms result to be more difficult than the *Internal Medicine* documents and only slightly easier than the *Medical Services* ones. The situation looks quite different if we consider instead the *Lexical* and the *Syntax* models: we can observe that the *Prevention* documents are easier–to–read than the documents of the other macro–specialties. On the contrary, sharp differences among the 4 macro–specialties and the 29 specialties occur as far as the *Lexical* and the *Syntax* models are concerned. In particular, all specialties turned out to be more difficult at the lexical than at the syntactic level. For what concerns the former, this follows from

the high percentage of out–of–vocabulary lemmas characterizing the informed consents with respect to the Basic Italian Vocabulary: as expected, *Prevention* documents represent an exception, being the easiest–to–read macro–specialty at the lexical level.

Consider now the results obtained with the *Syntax* model, according to which all informed consents turned out to be less difficult–to–read with respect to the lexical level. As discussed in Section 3.1, the typology of features contributing to this result is related to *local* aspects of the parse tree, taken in literature as an index of language complexity, rather than to *structural* complexity features. This type of evidence will be used in the near future to customize the set of features to be taken into account in the construction of a domain–specific version of the *Syntax* readability model. Also in this case, *Prevention* documents turned out to be more readable than the other specialties.

5 Conclusion and Future Work

In this paper, we illustrated the preliminary but encouraging results of a broader and long–term study devoted to enhance the accessibility of Italian health–related documents by relying on advanced Natural Language Processing techniques: the case study reported in the paper focuses on informed consent forms, which play a key role in doctor–patient communication. For this purpose, we used READ–IT, a general purpose NLP–based readability assessment tool for Italian. The results obtained so far show that the features automatically extracted from the linguistically annotated text and ranging across different levels of linguistic description, also including syntax, have a high discriminative power to guarantee a reliable readability assessment. To our knowledge, this is the first application of an advanced NLP–based methodology for readability assessment of Italian health–related documents. The proposed methodology was tested on a corpus of Italian informed consents currently used in healthcare trusts of the Regional Healthcare Service of Tuscany.

The results obtained by comparing readability scores across the considered medical specialties with respect to the different READ–IT models revealed that – generally speaking – informed consents are more difficult–to–read at the lexical level than at the syntactic level. This is in line with the linguistic profiling results discussed in Section 3.1, according to which the *2IC* corpus contains a higher percentage of out–of–vocabulary words, even higher than difficult–to–read texts (i.e. *Rep*). Behind this general trend, significant differences are reported for the different specialties, e.g. the *Prevention* documents turned out to be easier–to–read than the documents of the other (macro–)specialties.

The higher difficulty recorded at the lexical level suggests that the general purpose READ–IT tool needs to be specialized at the level of the permitted vocabulary, which should also include a selection of basic domain terms to be used in informed consent forms without any penalization at the level of the readability score. We are already working in this direction. Two experts in healthcare quality assessment are currently evaluating the out–of–vocabulary lemmas automatically extracted from the *2IC* corpus by the $T2K^2$ (Text–to–Knowledge) platform (Dell'Orletta et al., 2014c) with the final aim of creating a domain–specific lexicon to be used in the specialized version of READ–IT we are currently developing. The lexicon will be internally organized into three classes of *i)* "domain–specific words", i.e. words that cannot be avoided within health–related documents (e.g. *anestesia* 'anesthesia'), *ii)* "technical words", i.e. words that are specific to the domain but that should be explained with a gloss in order to be fully understood by laymen (e.g. *complicanza* 'complication'), and *iii)* "technicalities", i.e. words that are used by experts but that should be replaced with a simpler synonym in order to be fully understood by laymen (e.g. *fistola* 'fistula'). Obviously, as suggested above the specialization will also be concerned with grammatical features.

From a more general perspective, these preliminary results show a severe lack of knowledge and skills on the design of readable informed consents within healthcare services. Clearly, we can interpret these findings in the bureaucratic framework within which the documents are produced, missing the goal of informing patients while accomplishing the legal duty to have a "piece of paper" reporting the signatures of doctors and patient in the healthcare record, without a clear explanation of the treatments. Further research is needed to design and evaluate systems to support the preparation of the documents of informed consent: in this context, the customization of the READ–IT

Medical Specialty	n° documents	n° tokens	READ–IT		
			Base	Lexical	Syntax
Anesthesiology	20	21,065	50	93.37	69.62
Colorectal surgery	2	1,997	75.18	100	93.81
Obesity surgery	3	8,091	51.63	93.42	59.20
General surgery	19	11,588	43.03	78.29	58
Plastic surgery	4	3,550	88.95	98.72	96.51
Thoracic surgery	9	5,608	94.98	99.94	95.55
Vascular surgery	16	22,739	88.64	98.13	97.62
Ophthalmology	7	10,496	49.21	98.89	61.29
Otorhinolaryngology	134	194,421	25.14	94.90	69.42
Orthopaedics	44	76,712	50.54	97.58	89.66
Obstetrics and gynecology	35	31,243	60.37	97.31	58.52
Urology	17	19,576	85.40	98.08	89.16
TOTAL: Surgery	**313**	**407,086**	**63.59**	**95.72**	**78.19**
Cardiology	54	39,887	66.20	94.50	78.99
Diabetology	1	297	23.05	100	45.68
Gastroenterology	9	9,856	41.12	87.90	59.82
Neurology	8	5,199	69.44	97.96	94.98
Oncology	3	1,692	46.34	99.73	96.07
Pulmonology	4	3,220	49.57	98.18	78.27
Senology	17	20,455	85.09	99.68	93.88
TOTAL: Internal Medicine	**96**	**80,309**	**54.26**	**96.85**	**78.24**
Psychology	13	11,651	80.44	96.25	98.32
Screening	8	2,007	53.13	65.14	50.60
Vaccine	1	2,852	33.72	100	71.76
TOTAL: Prevention	**22**	**16,510**	**55.76**	**87.13**	**73.56**
Genetics	11	6,416	56.26	95.65	81.45
Immunohematology and transfusion	43	45,962	56.84	93.39	83.47
Nuclear medicine	29	18,045	52.62	96.56	68.48
Radiology	24	17,358	63.78	98.61	78.68
TOTAL: Medical Services	**107**	**87,781**	**57.38**	**96.05**	**78.02**
General	33	8,928	51.59	87.81	88.27
Pediatrics	13	6,092	49.84	99.46	74.67
Rehabilitation	2	674	63.84	99.99	96.25

Table 3: Readability assessment results by the *Base*, *Lexical* and *Syntax* models organized by medical specialties.

tool will play a key role. A specialized version of READ–IT will be possibly integrated within the Electronic Patient Record, so that the informed consent becomes part of a process of shared decision making where the doctors prepare a readable message for the patient at the time of the decision for a clinical procedure and collect questions and comments, that in turn feebacks into a software capable to learn from the daily practice. A limitation of this approach is the exclusive reliance on written documents, while according to the current debate (Korenman, 2015) in ethics and medico-legal issues the informed consent should be the result of a process of communication where the written document supports the doctor–patient communication. Bringing this to an extreme perspective, the informed consent could be simply the transcription of the dialogue that demonstrates the provision of comprehensive information on the possible treatments for a disease and the shared decision on the best alternative for the involved parts. However, even in this futuristic scenario NLP technologies could play a role.

References

S. Albolino, T. Bellandi, R. Tartaglia, and A. Biggeri. 2013. The incidence of adverse events in tuscany: results from a regional study involving 36 hospitals. *Proceedings of ISQUA 30th International Conference*, 13–16 October, Edinburgh.

E. Abrahamsson, T. Forni, M. Skeppstedt, and M.Kvist. 2014. Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2014, EACL Workshop)*, 57–65.

G. Attardi. 2006. Experiments with a multilanguage non-projective dependency parser. *Proceed-*

ings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X '06), 166–170.

G. Barlacchi and S. Tonelli. 2013. ERNESTA: A Sentence Simplification Tool for Children's Stories in Italian. *Proceedings of the 14th Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2013)*, 476–487.

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

D. Biber and S. Conrad. 2009. Genre, Register, Style. Cambridge: CUP.

T. Brants and A. Franz. 2006. *Web 1T 5–gram Version 1*. Linguistic Data Consortium, Philadelphia.

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

R. Clerehan R. Buchbinder, and J. Moodie. 2005. A linguistic framework for assessing the quality of written patient information: Its use in assessing methotrexate information for rheumatoid arthritis. *Health Education Research*, 20(3):334–344.

S. Cocchi, M. Mazzocut, C. Cipolat Mis, I. Truccolo, E. Cervi, R. Iori, and D. Orlandini. 2014. ETHIC. Evaluation Tool of Health Information for Consumers. Development, features and validation. *Divided we fall, united we inform. Building alliances for a new European cooperation, 14th EAHIL Annual Conference*, Roma (Italy), 11-13 June.

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

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

F. Dell'Orletta, S. Montemagni, and G. Venturi. 2011 READ–IT: assessing readability of Italian texts with a view to text simplification. *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, Edinburgh, UK, 73–83.

F. Dell'Orletta, S. Montemagni, and G. Venturi. 2014a. Assessing document and sentence readability in less resourced languages and across textual genres. *Recent Advances in Automatic Readability Assessment and Text Simplification. Special issue of International Journal of Applied Linguistics*, 165:2, John Benjamins Publishing Company, 163–193.

F. Dell'Orletta, M. Wieling, A. Cimino, G. Venturi, and S. Montemagni. 2014b Assessing the Readability of Sentences: Which Corpora and Features? *Proceedings of 9th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2014)*, Baltimore, Maryland, USA, 163–173.

F. Dell'Orletta, G. Venturi, A. Cimino, S. Montemagni. 2014c. T2K: a System for Automatically Extracting and Organizing Knowledge from Texts. In *Proceedings of 9th Edition of International Conference on Language Resources and Evaluation (LREC 2014)*, pp. 2062–2070, 26-31 May, Reykjavik, Iceland.

T. François and Eleni Miltsakaki. 2012. Do NLP and machine learning improve traditional readability formulas? *Proceedings of the NAACL-HLT 2012 Workshop on Predicting and Improving Text Readability for target reader populations (PITR 2012)*, 49–57.

A.K. Jha, I. Larizgoitia, C. Audera-Lopez, N. Prasopa-Plaizier, H. Waters, and D. Bates. 2013. The global burden of unsafe medical care: analytic modelling of observational studies. *BMJ Quality and Safety*, 22(10):809–15.

S. Kandula, D. Curtis, and Q. Zeng-Treitler. 2010. A semantic and syntactic text simplification tool for health content. *Proceedings of the American Medical Informatics Association Annual Symposium*, United States:366–70.

S. Kandula and Q. Zeng-Treitler. 2008. Creating a Gold Standard for the Readability Measurement of Health Texts. *Proceedings of the American Medical Informatics Association Annual Symposium*, Washington, DC, USA, 353–357.

D. Kauchak, O. Mouradi, C. Pentoney, and G. Leroy. 2014. Text simplification tools: Using machine learning to discover features that identify difficult text. *Proceedings of the 47th Hawaii International Conference on System Sciences (HICSS)*, Waikaloa, Big Island, Hawaii:2616–2625.

H. Kim, S. Goryachev, G. Rosemblat, A. Browne, A. Keselman, and Q. Zeng-Treitler. 2007. Beyond Surface Characteristics: A New Health Text-Specific Readability Measurement. *Proceedings of the American Medical Informatics Association Annual Symposium*, 418–422.

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

Korenman S. 2015. *Enduring and emerging challenges of informed consent*. N Engl J Med. 2015 May 28;372(22):2171-2.

M. Kvist and S. Velupillai. 2013. Professional Language in Swedish Radiology Reports – Characterization for Patient-Adapted Text Simplification. *Proceedings of the Scandinavian Conference on Health Informatics 2013*, Copenhagen, Denmark:55–59.

L. T. Kohn, J. M. Corrigan, D. S. Donaldson. 2000. *To err is human: building a safer health system*. Washington, DC: National Academy Press.

2015. Enduring and emerging challenges of informed consent. *New England Journal of Medicine*, 372(22):2171-2.

D. Gemoets, G. Rosemblat, T. Tse, and R. Logan. 2004. Assessing Readability of Consumer Health Information: An Exploratory Study. *Medinfo*, 868–873.

G. Grigonytė, M. Kvist, S. Velupillai, and M. Wirén. 2014. Improving Readability of Swedish Electronic Health Records through Lexical Simplification: First Results. *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2014, EACL Workshop)*, 74–83.

G. Leroy, E. Eryilmaz, and B.T. Laroya. 2007. Health information text characteristics. *Proceedings of the American Medical Informatics Association Annual Symposium*, Washington DC:479–483.

G. Leroy and J. E. Endicott. 2012. Combining NLP with Evidence-based Methods to Find Text Metrics Related to Perceived and Actual Text Difficulty. *Proceedings of the 2Nd ACM SIGHIT International Health Informatics Symposium*, Miami, Florida, USA:749–754.

G. Leroy, J. E Endicott, O. Mouradi, D. Kauchak, and M. L. Just. 2012. Improving Perceived and Actual Text Difficulty for Health Information Consumers using Semi-Automated Methods. *Proceedings of the American Medical Informatics Association Annual Symposium*, 522–531.

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

M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. 1999. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2), MIT Press:313-330.

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

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

J. Proulx, S. Kandula, B. Hill, and Q. Zeng-Treitler. 2013. Creating Consumer Friendly Health Content: Implementing and Testing a Readability Diagnosis and Enhancement Tool. *Proceedings of the 46th Hawaii International International Conference on-Systems Science (HICSS- 46 2013)*, 2445–2453.

A. Siddharthan. 2006. Syntactic Simplification and Text Cohesion. *Research on Language and Computation*, Volume 4, Issue 1, Springer Science, the Netherlands:77–109.

R. Tartaglia, S. Albolino, T. Bellandi, E. Bianchini, A. Biggeri, G. Fabbro, L. Bevilacqua, A. Dell'erba, G. Privitera, and L. Sommella. 2012. Eventi avversi e conseguenze prevenibili: studio retrospettivo in cinque grandi ospedali italiani. *Epidemiologia & Prevenzione*, 36(3-4):151–61.

G. Terranova, M. Ferro, C. Carpeggiani, V. Recchia, L. Braga, R. Semelka, and E. Picano. 2012. Low Quality and Lack of Clarity of Current Informed Consent Forms in Cardiology - How to Improve Them. *Journal of the American College of Cardiology (JACC): Cardiovascular Imaging*, Elsevier inc., vol. 5(6):649–655.

World Health Organization. 2015. WHO global strategy on people–centred and integrated health services. Interim Report available at `http://apps.who.int/iris/bitstream/10665/155002/1/WHO_HIS_SDS_2015.6_eng.pdf`

Q. Zeng-Treitler, H. Kim, S. Goryachev, A. Keselman, L. Slaughter, and C. Smith. 2007b. Text characteristics of clinical reports and their implications for the readability of personal health records. *Studies in Health Technology and Informatics*, 129(2):1117–1121.

Q. Zeng-Treitler, S. Goryachev, H. Kim, A. Keselman, and D. Rosendale. 2007a. Making Texts in Electronic Health Records Comprehensible to Consumers: A Prototype Translator. *Proceedings of the American Medical Informatics Association Annual Symposium*, 846–850.

Q. Zeng-Treitler, S. Kandula, H. Kim, and B. Hill. 2012. A Method to Estimate Readibility of Health Content. *Proceedings of the ACM SIGKDD Workshop on Health Informatics (HI-KDD 2012)*, Beijing, China.