ALTE 6th

INTERNATIONAL CONFERENCE

BOLOGNA, ITALY | 2017

Learning and Assessment: Making the Connections

3–5 May 2017

CONFERENCE PROCEEDINGS

ALTE
Association of Language Testers in Europe

In collaboration with

cliQ
CERTIFICAZIONE LINGUA ITALIANA DI QUALITÀ

and supported by

CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

Association of Language Testers in Europe, 1 Hills Road, Cambridge CB1 2EU, United Kingdom

www.alte.org

If you would like to quote any part of this work, please use the following details:

ALTE (2017). Learning and Assessment: Making the Connections – Proceedings of the ALTE 6th International Conference, 3-5 May 2017.

The opinions expressed in this work are those of the authors and do not necessarily reflect the views of ALTE. ALTE has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication, and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Learning and Assessment: Making the Connections

ALTE 6th International Conference, 3-5 May 2017

CONFERENCE PROCEEDINGS

Editor/Production Lead

Esther Gutiérrez Eugenio, ALTE Secretariat Manager

Production Team

John Savage, Publications Assistant, Cambridge English Language Assessment

Mariangela Marulli, ALTE Secretariat Coordinator

Alison French, Consultant

Acknowledgements

We would like to extend our most sincere thanks to all the contributors who have diligently replied to all our queries (even during the summer months) and thus allowed this publication to come to life on such a tight schedule.

Contents

Learning and Assessment: Making the Connections

Language learning, teaching and assessment…

… in a globalised economy

Predicting Readability of Texts for Italian L2 Students: A Preliminary Study

Giuliana Grego Bolli, CVCL (Centre for Language Assessment and Certification) – University for Foreigners of Perugia, Italy
Danilo Rini, CVCL (Centre for Language Assessment and Certification) – University for Foreigners of Perugia, Italy
Stefania Spina, Department of Human and Social Sciences – University for Foreigners of Perugia, Italy

Abstract: Text selection and comparability for L2 students to read and comprehend are central concerns both for teaching and assessment purposes.Compared to subjective selection. quantitative approaches provide more objective information, analysing texts at language and discourse level (Khalifa & Weir, 2009). Readability formulae such as the Flesch Reading Ease, the Flesch-Kincaid Grade Level and, for Italian, the GulpEase index (Lucisano and Piemontese, 1988), do not fully addressed the issue of text complexity. A new readability formula called Coh-Metrix was proposed (Crossley, Geenfield, & McNamara 2008), which takes into account a wider set of language and discourse features. A similar approach was proposed to assess readability of Italian texts through a tool called READ-IT (Dell'Orletta, Montemagni, & Venturi 2011). While READ-IT was tested on newspaper texts randomly selected, this contribution focuses on the development of a similar computational tool applied on texts specifically selected in the context of assessing Italian as L2. Two text corpora have been collected from the CELI (Certificates of Italian Language) item bank at B2 and C2 level. Statistical differences in the occurrence of a set of linguistic and discursive features have been analysed according to four different categories: length features, lexical features, morpho-syntactic features, and discursive features.

1 Introduction

The selection and the level of difficulty of texts to read is one of the central issues both for teachers and language testers. In the context of assessment, the decision taken with regard to texts in the case of reading tests has serious implication in the interpretation of test scores, hence in providing validity evidence to the overall testing process.

Focusing on reading comprehension, texts are mostly subjectively selected by experienced teachers and test producers depending on several aspects: specific curricula, programmes, guidelines and test specifications.

Other aspects such as the definition of readers' population, their linguistic needs, their educational background, their age, consequently involving other aspects such as text genre, text type, tasks to be assigned, are also taken into account.

There is quite a wide consensus in the literature about a set of other characteristics that can have an impact on the difficulty of a reading comprehension test, also in terms of cognitive demands imposed upon the reader (Bachman & Palmer, 2010; Purpura, 2014). These characteristics can be also measurable or judged by competent teachers or test developers, as often happens. They are: text length, grammatical complexity, word frequency, cohesion, rhetorical organisation, genre, text abstractness, subject knowledge and cultural knowledge. All these aspects relates to readability, which means to find measures of text's ease or difficulty in terms of comprehension (Green, Ünaldi & Weir, 2010; Khalifa & Weir, 2009).

Both qualitative and quantitative analysis can strongly contribute to a more comprehensive, evidence based approach to readability and hence on selecting and scaling texts in terms of difficulty both for assessment and teaching purposes. This kind of support is not provided by the Common European Framework of Reference for Languages (CEFR) descriptors and scales related to reading comprehension: they can provide information supporting text selection, but not in terms of readability.

Also within the CELI (Certificates of Italian Language) certification system, produced by CVCL (Centre for Language Assessment and Certification) at the University for Foreigners of Perugia, texts selection has been so far based on this set of characteristics subjectively assessed by CVCL experts' informed analysis.

Bearing in mind that in language testing terms, the decision taken with regard to texts in the case of reading tests may affect the interpretability of score outcomes, it is unquestionable that quantitative approaches, supported by automated analysis and systematic data collection, can provide more objective information, analysing texts on multiple levels of language and discourse and providing test producers and item writers ways to evaluate this aspect of test validity.

It is well known that readability assessment has been a central research topic for the past 80 years. The development of quantitative tools, such as Flesch Reading Ease, the Flesch-Kincaid Grade Level and, for Italian, the GulpEase index (Lucisano & Piemontese, 1988), opened the way to an automated textual description providing a more evidence-based approach to text selection and scaling.

Over the last decades, the automatic assessment of readability has received increasing attention: advances in computational linguistics and development of corpora, jointly with the availability of sophisticated language technologies, allow the capuring of a wide variety of more and more complex linguistic features affecting the readability of a text.

More recently, particularly in the last 20 years, scientific investigation of reading also benefited from more complex and automated measures of text characteristics, and systematic data collection, such as Coh-Metrix (Graesser, McNamara & Kulikowich, 2011; Graesser, McNamara, Louwerse & Cai, 2004) were proposed.

Taking all this into account, this paper reports on the development, at the University for Foreigners of Perugia, of a similar computational tool applied to texts specifically selected in the context of the CELI examinations suite. The tool itself and the consequent data collection and analysis will give more information and evidence about text readability as a part of the continuous validation process applied in the context of CELI.

2 Selection of texts and tasks in CELI 3 (B2) and in CELI 5 (C2)

In the routine work of texts selection for the Reading component of CELI exams, the following basic aspects are taken into account by CVCL item writers: the characteristics of texts as shown in the CEFR, and the genres identified in the Profilo della Lingua Italiana (Spinelli & Parizzi, 2008). A detailed overview of CELI exams specifications can be found in Grego Bolli & Spiti (2004).

Amongst the CEFR descriptors concerning Reading skills used in the selection of texts, “Overall reading comprehension”, “Reading for orientation” and “Reading for information and argument” can be found. Along with them, the Profilo helps in identifying the genres and text types. It has to be underlined how the Profilo does not include any referential for C2 level, but, on the other hand, C2 level language users can deal with any type of textual genre. With these indications in mind, in CELI exams the text types used for assessment of language competence for CELI 3 (B2) include fiction and non-fiction books, magazine and newspaper articles, textbooks, interviews, and personal letters, whereas for CELI 5 (C2), fiction and non-fiction books, including literary journals, specialist magazines, newspapers, textbooks and essays, personal letters, regulations, memoranda, reports and papers are used.

Tasks in the exam papers have the objective of testing the following sub-skills, for CELI 3: reading for gist; identifying point of view; identifying main points; reading for detailed information, skimming and scanning; and for CELI 5, along with the above mentioned: identifying point of view and tone, guessing meaning from context, recognising the organization of a text, reading for detailed information. The length of texts used for testing reading skills vary from 250–350 words in CELI 3 to 600–650 in CELI 5, and in both papers the answer format for the texts taken into account include 4-option multiple-choice, and short answers. It has to be added that items are generally calibrated according to IRT model based on Rasch analysis, placing the item difficulty at the pre-established level.

3 Method

The main question we are trying to answer in this study is: how can we operationalise complexity in order to measure it in texts to be selected for learning and assessment purposes?

From the theoretical point of view, we considered two different models in the field of measures of complexity.

The first model is Coh-Metrix (Graesser et al., 2004): it is a well-established project that takes into account a wide set of language and discourse features, based on 108 indices. While these indices belong to different levels of linguistic analysis, Coh-Metrix is mainly focused on cohesion, and is specifically targeted to English texts.

The second model is READ-IT (Dell’Orletta et al., 2011), which is targeted to Italian texts, and aimed at text simplification: its intended audience are mainly people with low literacy skills and/or with cognitive impairment. In contrast with Coh-Metrix, READ-IT is mainly focused on lexical and syntactic features, such as syntactic dependencies or part-of-speech probability.

None of these two models is sufficient to achieve the goal of developing a computational tool to be used with texts specifically selected in the context of learning and assessing Italian as an L2. The methodology we followed was based on two different steps: the corpus-based feature selection process, and the tool creation and testing.

The first task we had to perform was the identification of a set of linguistic features to be used in order to establish text difficulty. As we still are at an early stage of the project, in the features selection process we preferred easy-to-identify features which could be reliably detected within the output of computational resources.

To this aim, we collected two corpora of texts from the CELI item bank at B2 and C2 level. It is important to stress that these texts were selected and assigned to a specific level by experienced, professional teachers. With the 213 selected texts, we built a corpus with 133,364 tokens (B2 level: 122 texts and 59,423 tokens; C2 level: 91 texts and 73,941 tokens), which was xml-annotated and post-tagged (the tag-set and annotation scheme were the same as those used for the annotation of a reference Italian corpus; see Spina, 2014).

As complexity is intrinsically multifactorial, we selected a wide set of linguistic and discursive features, that, in our opinion, affect texts comprehension and systematically vary as a function of types of texts and grade level. In addition, these features show a growing computational complexity, so as to follow the different levels of linguistic analysis automatically carried out on texts.

The selected linguistic features are distributed in the following four categories:

- raw-text features (length features)
- lexical features
- morpho-syntactic features
- discursive features.

4 Results

4.1 Raw-text features

Raw-text features are from the computational point of view the simplest category, and were typically used within traditional readability metrics. Nevertheless, they can give a contribution in predicting text complexity: higher level texts (C2) are formed by longer sentences (B2: 18.1; C2: 20.8 words per sentence), and by slightly longer words (B2: 4.8; C2: 5 mean word length).

4.2 Lexical features

We selected four different lexical matrix that are generally considered in the computation of linguistic complexity: lexical diversity (Aluisio, Specia, Gasperin, & Scarton, 2010), lexical density (Feng, Elhadad & Huenerfauth, 2009), basic Italian vocabulary rate (Dell'Orletta et al., 2011), and the percentage of concrete/abstract nouns.

Lexical diversity (Malvern, Richards, Chipere & Durán, 2004), defined as the ratio of total number of words to the number of different unique words, is a measure of the amount of different words used in a text. A text with a higher score of lexical diversity includes more different words, and is therefore more complex, while texts with lower scores tend to repeat the same words many times. We used the Guiraud index (Guiraud, 1954) as an index of lexical diversity. This index was used instead of type/token ratio because it compensates the systematic decrease of the number of tokens when texts to compare have different lengths (e.g. Van Hout & Vermeer, 2007). The respective values of lexical diversity (B2: 43.3; C2: 51.9) show that higher level texts tend to include more different words, and, as a consequence, to be more complex.

Lexical density (Ure, 1971) refers to the ratio of content words (verbs, nouns, adjectives and adverbs) to the total number of words in a text. The idea behind this measure is that more dense texts are also more difficult to understand. The respective values (B2: 44.7; C2: 45.8) show that lexical density also contribute to the greater complexity of C2 texts.

The basic Italian vocabulary rate measures the internal composition of the vocabulary of the texts. To this end, we took as a reference the New Basic Italian Vocabulary (NVdB) by De Mauro & Chiari (forthcoming), that includes a list of 7,500 words highly familiar to native speakers of Italian. In more detail, we calculated two different features corresponding to: a) the percentage of lemmas on this reference list that are used in the texts; b) the internal distribution of the occurring basic Italian vocabulary words, and in particular the 2,000 most frequent, or fundamental words. Both the 7,500 total words (B2: 77.9; C2: 76.4) and the 2,000 fundamental words (B2: 71.2; C2: 68.8) are used more in the easier texts. This reveals, hence, a greater use of more frequent, and thus easier, lexical items in lower level texts.

Finally, we considered the use of concrete and abstract nouns. The percentage of concrete nouns is significantly higher in B2 texts (B2: 56.7; C2: 48), while abstract nouns are used more in C2 texts (B2: 11.8; C2: 18.3). This finding is relevant for our research, because concrete nouns are more familiar and then easier for the reader, as familiarity has a strong impact on a wide range of cognitive processes, including comprehension.

4.3 Morpho-syntactic features

In general, the morpho-syntactic features selected for this study seemed to affect texts complexity less than other linguistic features: we did not find significant differences in part-of-speech distribution and in the global number of subordinate clauses, although subordination is traditionally acknowledged as an index of structural linguistic complexity. The only kind of subordinate clause that is used significantly more in C2 texts is relative (log-likelihood = 13.40).

4.4 Discursive features

We believe that cohesion plays a key role in text readability. By cohesion we refer to the “characteristics of the explicit text that play some role in helping the reader mentally connect ideas in the text" (Graesser, McNamara, & Louwerse, 2003).

Following the Coh-metrix model, we studied two different dimensions of cohesion: the referential cohesion and the deep cohesion.

Referential cohesion can be measured by assessing the overlap between adjacent sentences: high cohesion texts contain words that overlap across sentences, forming threads that help readers to recover the message, while low cohesion texts have to count on knowledge-based inferences to fill the gaps.

What we found in our data was that adjacent sentences that contain overlapping nouns are significantly more frequent in B2, easier texts.

The following example shows the use of the overlapping noun medico ("physician") accross three adjacent sentences.

> Conosco medici laureati con 110 e lode da cui non mi farei curare nemmeno un'unghia. Ho fiducia in questo medico falso. Non lo cambierei con nessun altro medico.

Moving to deep cohesion, taking for granted that cohesion gaps increase reading time and complexity, we measured the use of connectives, which play an important role in the creation of logical relations within text meanings, and provide clues about text organisation (Halliday & Hasan, 1976).

Based on eight classes of connectives (causal, temporal, additive, adversative, marking results, transitions, alternative or reformulation/specification), we found that in some cases, as in causal connectives, there is a substantial equivalence in the two levels of texts, but in other cases, as in temporal connectives, there is a significant difference, and connectives are much more frequent in easier texts.

5 Conclusions

We presented an exploratory study on the possibility of measuring complexity in Italian texts, selected for L2 learning and assessment purposes.

The process of corpus-based feature selection, resulting in four dimensions with growing computational complexity, revealed significant differences in texts assigned to specific CEFR levels by experienced teachers. These differences emerged particularly in lexical and discursive features. This analysis also confirmed that the use of a quantitative approach should always be part of the cyclic process of text selection.

Future work will be needed in order to fulfil the aim of creating a tool for the automatic assessment of complexity. One future direction will be the refinement of the linguistic indices of complexity, with a deeper analysis of overlap across sentences, and the addition of narrativity, which is a major predictor of text complexity.

References

Aluisio, S., Specia, L., Gasperin, C., & Scarton, C. (2010). *Readability assessment for text simplification*. Paper presented at the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications.

Bachman, L. F. & Palmer, A. , S. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.

Crossley, S.A., Greenfield, J., & McNamara, D.S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly, 42*(3), 475–493.

De Mauro, T., & Chiari, I. (forthcoming). *Il Nuovo Vocabolario di Base della Lingua Italiana*.

Dell'Orletta, F., Montemagni, S., & Venturi, G. (2011). *READ–IT: Assessing readability of Italian texts with a view to text simplification*. Paper presented at the 2nd Workshop on Speech and Language Processing for Assistive Technologies, Edinburgh.

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

Green, A., Ünaldi, A., & Weir, C. J. (2010). Empiricism versus connoisseurship: establishing the appropriacy of text for testing reading for academic purposes. *Language Testing, 27*(3), 1–21.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. In A. P. Sweet & C.E. Snow (Eds.), *Rethinking reading comprehension* (pp. 82–98). New York: Guilford Publications.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, And Computers, 36*, 193–202.

Grego Bolli, G. & Spiti, M. G. (2004). *Misurare e valutare nella certificazione CELI*. Perugia: Edizioni Guerra.

Guiraud, P. (1954). *Les caractères statistiques du vocabulaire. Essai de méthodologie*. Paris: Presses Universitaires de France.

Halliday, M. A. K. & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Khalifa, H. & Weir, C. J. (2009). *Examining reading: research and practice in assessing second language reading*. Cambridge: UCLES/Cambridge University Press.

Lucisano, P. & Piemontese, M.E. (1988). GULPEASE: una formula per la predizione della difficoltà dei testi in lingua italiana. *Scuola e città, 31*(3), 110–124.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development. Quantification and assessment*. London: Palgrave Macmillan.

Purpura, J. E. (2014). Cognition and language assessment. In Kunnan, A., J. (Ed.). *The Companion to Language Assessment volume III* (pp. 1,453–1,476). Oxford: Wiley Blackwell.

Spina S. (2014). Il Perugia Corpus: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione. In R. Basili, A. Lenci, & B. Magnini (Eds.), *Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014* (354–359). Pisa: Pisa University Press.

Spinelli, B. & Parizzi, F. (2010). *Profilo della lingua italiana. Livelli di riferimento del QCER A1, A2, B1, B2*. Firenze: La Nuova Italia.

Ure, J. (1971), Lexical density and register differentiation. In G. Perren & J. L. M. Trim (Eds.), *Applications of Linguistics. Selected Papers of the Second World Congress of Applied Linguistics* (pp. 443–452). Cambridge: Cambridge University Press.

Van Hout, R. & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and Assessing Vocabulary Knowledge* (pp. 93–115). Cambridge: Cambridge University Press.

European Commission

Under the auspices of the Secretary General of the Council of Europe, Mr Thorbjørn Jagland

COUNCIL OF EUROPE

CONSEIL DE L'EUROPE