Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 198 (2015) 515 – 525

Procedia
Social and Behavioral Sciences

7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)

Building a Corpus of 2L English for Automatic Assessment: the CLEC Corpus

Mª Ángeles Zarco Tejada*, Carmen Noya Gallardo†,

Mª Carmen Merino Ferradá‡, Mª Isabel Calderón López§

Dpto. Filología Francesa e Inglesa, Universidad de Cádiz, Avda. Doctor Gómez Ulla, 1, 11003 Cádiz, Spain.

Abstract

In this paper we describe the CLEC corpus, an ongoing project set up at the University of Cádiz with the purpose of building up a large corpus of English as a 2L classified according to CEFR proficiency levels and formed to train statistical models for automatic proficiency assessment. The goal of this corpus is twofold: on the one hand it will be used as a data resource for the development of automatic text classification systems and, on the other, it has been used as a means of teaching innovation techniques.

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.

Keywords: automatic assessment; CEFR proficiency labels; teaching innovation techniques; corpus linguistics; automatic linguistic profile.

* Mª Ángeles Zarco Tejada. Tel.: +00-34-956015523; fax: +00-34-956015501.
E-mail address: angeles.zarco@uca.es

Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.
doi:10.1016/j.sbspro.2015.07.474

1. Introduction

Nowadays one of the main problems in our University, as far as granting our students with a language proficiency certificate, is concerned with the production of 2L English materials for language proficiency assessment. Students are to be provided with a proficiency level degree according to the levels described by Common European Framework of References for Languages (CEFR). But, as CEFR authors say, the CEFR is deliberately atheoretical (Council of Europe, 2001) and adopts an action-oriented approach, describing language learning outcomes in terms of language use. Since then, there have been many groups, projects and research activities dealing with language testing and second language acquisition across Europe. One of the main goals has been the identification of criterial features for L2 English for each CEFR level (Salamoura and Saville, 2010), basic aim of the Cefling project (Alanen, Huhta, and Tarnanen, 2010) or the English Profile project (Hendriks, 2008; and Kurtes and Saville, 2008), among others.

Thus, following Alanen, Huhta, and Tarnanen (2010) and Hulstijn, Alderson, and Schoonen (2010), and their insights on Second Language Acquisition (SLA) and language testing research, we have decided to collect data from existing language texts already classified according to CEFR levels and analyze them in terms of linguistic features (Banerjee, Franceschina, and Smith, 2004; Norris, 1996; Norris and Ortega, 2009).

As Dahlmeier, Ng, and Wu (2013) point out, the success of statistical methods in NLP over the last two decades can largely be attributed to the availability of large annotated corpora that can be used to train statistical models for various NLP tasks. In this sense, our ultimate goal in making this corpus is to provide a linguistic resource for automatic text classification following a similar approach carried out for linguistic profiling of texts in Italian by Montemagni (2013) and Dell'Orletta, Montemagni, and Venturi (2013).

So, our project was set up in 2012. We have developed CLEC (CEFR-Labeled English Corpus) with more than 200.000 words of grammatical English examples taken from 2L English texts already classified for the CEFR levels A1, A2, B1, B2, C1 and C2. The texts have been manually encoded and are divided in different groups corresponding to A1, A2, B1, B2, C1 and C2 CEFR levels. Our Corpus follows language-oriented criteria, not communicative criteria, since the classified CEFR texts used have been labeled according to linguistic facts. The corpus has been annotated with additional information as metadata, so that each text has an identification mark, a reference to the main grammatical structures and a reference to the main language function identified in the text.

The creation of this corpus has been used for teaching innovation performance as well, since students of the English Studies grade have been involved in its construction not only collecting material activities but also encoding sentences and annotating texts.

In this paper we describe the corpus in detail. We give a short introduction to the background and goal of our project and provide a full description of the process of building CLEC, the corpora used, the annotation scheme and the problems arisen.

2. Background

The point of departure for an adventure such as developing CLEC is closely related to practical needs emerged in our University with the duty of granting our students with an English proficiency certificate and, thus, the complicated task of producing 2L English materials for language proficiency assessment. In this sense, having in mind the large amount of texts to be prepared since the demand for English proficiency certificates was increasingly requested, we decided to build up a corpus of CEFR-labeled English texts to be used as a linguistic resource for automatic text classification, following a similar approach launched by Montemagni (2013), Dell'Orletta and Montemagni (2012), Dell'Orletta, Montemagni, and Vecchi (2011), Dell'Orletta, Montemagni, and Venturi (2011, 2012, 2013) for Italian texts. As Montemagni (2013: 20) points out: "… identified monitoring parameters […] can

be usefully employed for monitoring linguistic competence of L1 and L2 learners of Italian". Within the recent trend of using NLP techniques to study linguistic form instead of content of a text and following Heilman, Collins-Thompson, Callan, and Eskenazi (2007) and Collins-Thompson and Callan (2005) approach on NLP uses for L1 and L2 text readability measuring, these authors (Montemagni, 2013; Dell'Orletta, Montemagni, and Vecchi, 2011; Dell'Orletta, Montemagni, and Venturi, 2011, 2012, 2013) show how to classify texts according to their genre or readability levels by the automatic identification of linguistic features. For this task, as they explain, they use READ-IT that, given a set of features and a training corpus, creates a statistical model used for assessing the readability of new texts. In this sense, our corpus organized by levels of proficiency will act as a "trainer" and will provide texts already classified by levels of proficiency helping the system to identify linguistic features for each level. As Dahlmeier, Ng, and Wu (2013: 22) point out: "… The success of statistical methods in NLP over the last two decades can largely be attributed to advances in machine learning and the availability of large, annotated corpora that can be used to train and evaluate statistical models for various NLP tasks".

In fact, in the last decade computational linguistic technologies have been applied for assessing linguistic competence with different purposes such as deficit cognitive analysis through syntax procedures (Roark, Mitchell, and Hollingshead, 2007), development of child language via complex syntax use (Sagae, Lavie, and MacWhinney, 2005), text readability measuring with the ranking of documents by reading difficulty as one of their applications, as mentioned above, or reading abilities as a component of linguistic proficiency (Petersen and Ostendorf, 2009).

As READ-IT did for Italian and focusing on lexical and syntactic features, proficiency assessment will be a classification task: given a set of texts classified from A1 to C2 CEFR levels, the system will be able to discern among levels and identify proficiency features of new texts classifying them with a label.

As computational linguistic technologies have been increasingly used within the teaching sphere, the number of corpus is progressively growing to achieve different goals. Most of them, though, are either collection of texts produced by learner students of a second language: The Cambridge Learner Corpus (CLC), the Cambridge English Profile Corpus (CEPC), the NUS Corpus of Learner English, CEFLING, etc., or a collection of texts that represent either written or oral language as a first language (BOB, LOB, BNC, or C-ORAL-ROM Italia for Italian), to give just a hint. In fact, the main obstacle for automatic assessment of text according to CEFR classification is the absence of corpora already classified. Our goal then is to produce CLEC, a CEFR-labeled English Corpus for automatic proficiency classification of texts.

A completely different approach is put forward in The Profile Program, a collaborative programme endorsed by the Council of Europe, designed to create a set of reference level descriptions for English. One of the main aims is to provide examples of the competences laid out in the CEFR by supplying grammar and vocabulary examples as well as function descriptions and, thus, becoming a benchmark for English proficiency at each level of the CEFR. As they explain, the English Profile Programme sets up as the latest phase of a process that started with the Threshold series (van Ek and Trim, 1989a, 1989b, 2001) during the 80s and tends to be a reference for the production of course materials, teachers, teaching guides, words lists and any sphere having to do with language learning. The examples being used to describe English competences are examples produced by learners of English, so, having an empirical methodology. As a source, The English Profile Programme is certainly an important corpus resource since the CEPC aims to collect 10 million words of spoken and written language and covers from A1 to C2 CEFR levels. So far it is an on-going project and it is not available unless you get involved as a researcher.

3. The corpus

3.1 Data collection

The project started in October 2012 and is still going on. In the beginning the corpus was called Eng-Corpus and every year it has been financed by the Teaching Innovation Section of the University of Cádiz to address several meetings with researchers of the Istituto di Linguistica Computazionale (Pisa). Four teachers of the Department, authors of this article, are in charge of the project with several tasks to accomplish, among others, to carry out the main corpus designing tasks, the classification of materials, the organization of students, the making of the e-platform "Corpus", the revision of exercises and the codification of texts. Besides the teachers we counted on about 10 collaborating students, a student of our PhD program and a post-graduate student, that were very much involved in the codifying process and that participated with us in the several meetings where main difficulties were discussed. These students were all doing the English Studies degree; students of 2nd, 3rd and 4th year had a level of English equivalent to a B2-C1, whereas the post-graduate ones had a C2 level. Finally, about 30 undergraduate students of Syntax, Discourse Analysis and Computational Linguistics subjects of the English Studies grade were willingly involved in this project to collaborate in the codification of texts through the making of grammatical exercises.

The data collection was distributed each year as follows:

Year 2012-13: From February 2013 to June 2013 our corpus had an amount of 60723 words distributed in the following CEFR levels:

- A1: 3744 words
- A2: 20322 words
- B1: 35383 words
- B2: 1274 words

Year 2013-14: From December 2013 to June 2014 we had an amount of 105949 words distributed in the following levels:

- A1: 3744 words
- A2: 21239 words
- B1 45864 words
- B2: 11189 words
- C1: 3648 words
- C2: 20265 words

Year: 2014-2015: We started in December and our work is going on. Our main focus this year is to include listening exercises of oral speech. We are mainly concerned with having texts that show oral English. The total amount of words is 237958. The distribution of data is as follows:

- A1: 3744 words
- A2: 21239 words
- B1 79923 words
- B2: 48088 words
- C1: 64699 words
- C2: 20265 words

The basic statistics of CLEC are shown in Table 1:

Table 1. Basic statistics of CLEC in January 2015

CEFR levels	A1	A2	B1	B2	C1	C2
Number of words	3744	21239	79923	48088	64699	20265
Number of files	62	160	210	162	156	148
Number of written English texts	62	160	174	63	37	148
Number of oral English Texts	-	-	36	99	119	-

3.2 Codifying process

Our corpus codifying process was articulated as an optional activity of the Teaching Innovation Program held at the University of Cádiz. All students of the English Studies diploma and some others doing other Philological studies (Spanish, French, Linguistics, Classical studies) were informed of this research project we challenged to attempt. The codifying process was set up, thus, as an extra activity for those students that wanted to participate. The innovative thing here was that students were informed of the research project with an exhaustive explanation on the method of codification, the amount of data to achieve and the goal to reach. They were aware of being part of a research process, something new in our department. As explained in the Teaching Innovation Program, we bring scientific research closer to the students in order to improve their University academic training. The fact that this project was born in collaboration with ILC-Pisa gave it an especial nature making it more attractive, if possible.

Within the students, a distinguished role had all the "Collaborating students", students awarded with a scholarship to collaborate with teachers in different academic tasks, and students within the PhD program, who had the prominent task of organizing materials and codifying texts.

CLEC consists of about 200000 words distributed in classified texts by levels of proficiency. The data source is the set of books of 2L English materials used for teaching activities at the English department. The teaching materials used were the pre-intermediate to advance set of the New Headway, New English File and Face2face student's books. All text examples have been done by our students as homework activities to test their linguistic proficiency. In a second phase, results have been checked for grammatical errors and have been corrected. As a result of this process, we have collected a group of 898 files with texts classified according to CEFR levels and annotated with grammatical and functional information. Students were tutored on the codification procedure, so that, they were informed of the need to save each text in a different file, to save it as a plain text and load it in the e-platform "Corpus", an e-learning tool that the University of Cádiz allowed us to have, as well as to follow some very general instructions on the codification of metadata.

The platform used as a repository of our research project has six main areas corresponding to each CEFR level; that is to say, there is an A1, A2, B1, B2, C1 and C2 sections. Students willing to participate in the creation of CLEC were distributed in levels having in mind their levels of proficiency. Students of the first and second year of English Studies and other Philological Studies were placed to make exercises of the lower levels, A1 and A2, whereas students of the upper levels of English Studies and students of the PhD Programme dealt with exercises of levels B1, B2, C1 and C2. The platform allows students to upload homework assignments so each student has its place with a task link to send his/her texts.

3.3 Structure of texts

Every text has two tabs, the opening tab at the beginning of the text and the closure one at the end. The main information is included in the beginning tab as metadata where three elements are compulsory for each text:

- Id (identification), where we include the source of the text. Elements of information such as the CEFR level (A1, A2… C2), the student's book names and the unit are part of the id argument. The student's name is not part of the metadata information of each text although in the e-platform "Corpus" all texts are classified according to CEFR levels and name of the student, so we have a trace of the student in charge of codifying each exercise.

- Cat (category), where the main linguistic function is mentioned. This argument is one of the arguments we had more problems with and, in fact, we think is one of the elements to be improved since there are some inconsistencies along the whole Corpus. As we have already mentioned, CEFR (Council of Europe, 2001) describes language learning outcomes in terms of language use, thus, adopting an action-oriented approach. For this reason, we thought it as useful information to have our texts classified for linguistic functions so that we could analyze them afterwards within a language use perspective. In this sense we made use of the Waystage (1998), Threshold (1998) and Vantage (2001) specifications of learning objectives developed within the Council of Europe's Programme for language learning in Europe, and included as part of the goals of the CLEC Corpus the analysis of the main linguistic function a text was a representation of. This way we included as part of the metadata of some part of the texts the linguistic function conveyed. We faced different problems when dealing with linguistic functions and texts. The main problem, though, lies in the fact that most texts represent more than one linguistic function, so it is hard to determine just one or even the main one. As the Council of Europe Programme mentions, the language functions specified are: imparting and seeking factual information, expressing and finding out attitudes, suasion, socializing, structuring discourse and communication repair.

- Arg (argument) where the main grammatical task represented by the text is given. In this argument we include grammatical information. Since most exercises are classified in units of different grammatical content, the grammatical information provided by our sources guided the grammatical data indicated as metadata for each text.

These are some examples of the opening and end tabs used at the beginning of the codified texts:

```
<doc id=”B1 NH Intermediate Unit 9” cat=”Expressing and finding out attitudes” arg=”conditionals”>
(…text...)
</doc>

<doc id=”B1 NH Pre-Intermediate Student's book. U4” cat=”Imparting and seeking factual information” arg=”Articles”>
(…text...)
</doc>

<doc id="A2 NH Elementary Unit4" cat="Socialising" arg="questions and answers">
(…text...)
</doc>
```

Figure1 below gives an example of the Keyword in Context" (KWIC) conditional sentences encoded in B1 level of the CLEC and tagged by AntConc3.4.3:

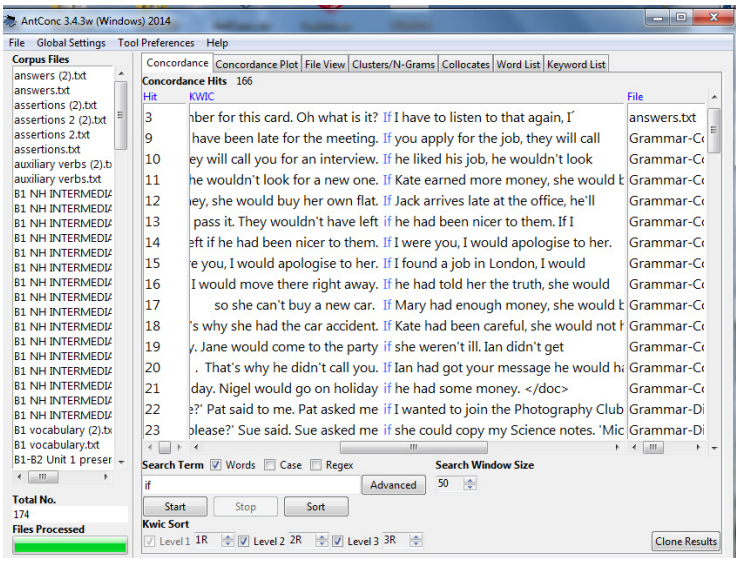

Fig. 1. A B1 example of KWIC conditional sentences

3.4 Main obstacles in the CLEC construction

The CLEC corpus covers the 6 levels specified by CEFR but we have especially focused on those ones our students would more likely submit language proficiency exams, that is to say, B1, B2 and C1 levels. In fact, the vast amount of words in these levels contrasts with the number of words found in A1, A2 or C2. This unbalanced collection is a matter of concern since comparative studies among levels regarding linguistic facts may thus have inaccurate results.

On the other hand, one of the most hard-working tasks has been supervising the student's assignments. Teachers and post-graduate students spent a lot of time checking the opening tabs and spelling or grammatical errors in the texts. Among these, as mentioned above, the most time-consuming activity was revising the linguistic function encoded in the argument "cat" of the opening tab since not all students managed to identify it and, if they did, sometimes there were inconsistencies. It still remains a difficulty to adjust.

Finally, this year we have dealt with oral English examples. Typing examples of oral English was part of listening exercises uploaded in the e-platform. Again correcting these texts was tedious and monotonous.

4. Linguistic profiling of CLEC: first results

The linguistic profiling of texts is a first step towards our ultimate aim of producing automatic proficiency assessment of new texts. In the meanwhile the system is developed, the linguistic analysis of our corpus can help to identify and define the criterial CEFR levels features. Thus, the results mentioned here are the first outcome applied to levels A2, B1 and B2 of written English and to B1 and B2 levels of oral English. The linguistic profiling of these levels follows the methodology and linguistic description explained in Montemagni (2013). Such an approach is based on the identification of the linguistic structure of texts through a multi-level linguistic analysis that includes the analysis of characters, words, morphological categories or syntactic structures. Accordingly, vast amounts of texts and computational linguistic techniques make it possible to analyze texts and identify significant linguistic features. Basically, the occurrences of the selected linguistic features are counted for the identification of the text

profile (Biber, 1988; van Halteren, 2004). The linguistic structure identification of the text is driven step by step starting by tokenization, where the text is divided in words, followed by a morphosyntactic analysis, where each token is assigned a POS tag and a dependency relation among words is established. After this phase of linguistic annotation, our corpus is ready for other types of automatic processing very useful for the linguistic profiling of texts. Differences among levels of proficiency are based on text readability complexity. So far, as far as linguistic complexity is concerned, either lexical or syntactic complexity is analyzed. Syntactic tree depth is considered a central aspect for text readability assessment (Yngve, 1960; Frazier 1985; Gibson 1998) as token-dependent distance is another aspect of readability measures (Lin, 1996; Gibson, 1998). Within the syntactic sphere linguistic complexity can be represented by the number of dependents of verbal syntactic categories, number of verbal heads and type of verbal valence in each sentence or number of subordinate clauses. The analysis of subordinate sentence types has not been carried out in this paper. Finally, lexical complexity is another factor that determines readability measures that we have formalized in terms of the number of tokens each sentence has, the number of characters within tokens, and the type/token ratio that reflects lexical variation in a corpus.

The number and types of features are deeply explained in Montemagni (2013). We introduce here very shortly just those we find remarkable for our study. These are displayed in table 2 below:

Lexical features:
Sentence length: average number of words per sentence.
Word length: average number of characters per token.
Type/Token ratio: it is calculated with respect to lemma. It measures the vocabulary richness of a corpus. Values are between 0 and 1. Figures closer to 0 indicate low lexical variation and those ones closer to 1 indicate high variation.

Morpho-syntactic features:
Verbal heads per sentence: average number of verbal occurrences in a sentence.
Verbal dependents: average of dependents of verbal heads.
Subordinate sentences: average number of subordinate sentences.
Link length: length of the dependency relation between head and dependent. The length is measured in terms of distance in tokens between head and dependent.
Tree depth: depth of the tree calculated in terms of the longest path from the root to some leaf.

Figures in table 2 are according to predictions: in the lexical area the average number of tokens, average number of characters per token and Type/Token ratio increase as the level of proficiency is higher. When coming across with syntactic structure, the average number of verbal heads increases from A2 to B2 and figures of token-dependent distance, tree depth, subordinate clauses or verbal valences 3 and 4 show a remarkable increasing difference showing a deeper level of structural linguistic complexity in higher levels of proficiency.

Table 2. Linguistic profiling of written A2, B1 and B2 of CLEC

Linguistic Text Features	A 2	B 1	B2
N. of token per sentence	7,571	9, 566	15,820
N. of characters per token	3,921	4,020	4,626
100 Type/Token	0,416	0,541	0,582
Verbal heads per sentence	1,216	1,658	2,011
N. of dependents per verbal heads	1,120	1,278	1,218
Token-dependent distance	2,810	3,631	6,352
Tree depth	2,729	3, 358	4,852
Subordinate clauses	17,796	19,427	23,716

Verbal valence 2	60,651	59,053	47,011
Verbal valence 3	20,990	26,309	26,017
Verbal valence 4	2,739	4,856	5,514

In table 3 we show the results obtained for B1 and B2 levels of oral English. These figures are according to predictions too: B2 level shows longer words, sentences with a greater number of words and with a richer vocabulary (Type/Token ratio), and syntactic complexity is higher than in B1 level as it is shown with all results regarding the number of verbal heads and verbal dependents per verbal head, complex Noun Phrases, token-dependent distance, tree depth, subordinate clauses or verbal valences 3 and 4.

Table 3. Linguistic profiling of oral B1 and B2 of CLEC

Linguistic Text Features	B 1	B2
N. of token per sentence	10,255	16,074
N. of characters per token	3,689	3,848
50 Type/Token	0,731	0,814
Verbal heads per sentence	1,578	2,310
N. of dependents per verbal heads	1,266	1,372
Nominal-dependent length	1,057	1,125
Token-dependent distance	1,864	2,274
Tree depth	3,251	4,547
Subordinate clauses	8,014	9,367
Verbal valence 2	59,051	52,407
Verbal valence 3	25,842	28,760
Verbal valence 4	4,608	6,993

Such an approach has been successfully used for the profiling of other text types and for other goals, such as the profiling of the Italian language within different diamesic, diastratic, and diaphasic varieties (Montemagni, 2013); the identification of similarities and differences of Italian learners L1 and L2 written texts and teaching materials at school (Dell'Orletta, Montemagni, and Vecchi, 2011); the profiling of writing improvement at school (Barbagli, Lucisano, Dell'Orletta, Montemagni, and Venturi, 2014). Besides, these features have been used for the automatic assessment of text readability (Dell'Orletta, Montemagni, and Venturi, 2011), for classifying documents according to text genre (Dell'Orletta, Montemagni, and Venturi, 2014) and for automatic identification of L1 from L2 production (Cimino, Dell'Orletta, Venturi, and Montemagni. 2013).

5. Conclusions

In this study we have described the CLEC corpus built up to train statistical models for automatic proficiency assessment. We have explained how we managed to develop this corpus as part of the innovation teaching techniques project set up at our university. The main problems we faced in the creation process dealt with the linguistic function specifications for each text. In fact, identifying linguistic functions were too complex sometimes, mainly for texts with more than a single linguistic function. The profiling results obtained for A2, B1 and B2 written texts and for B1 and B2 oral texts make evident that a readability assessment of our corpus is a first step towards the automatic identification of proficiency levels. As expected, either the lexical features or the syntactic ones show deeper levels of complexity in higher levels of proficiency. Future research includes: 1) the study of POS categories

in each level and in a compared written-oral English analysis; 2) the study of sentence dependence types; 3) the definition and organization of linguistic functions criteria for text classification.

Acknowledgements

We would like to thank the Italian Natural Language Processing Laboratory (ItaliaNLP Lab) of the Istituto di Linguistica Computazionale "Antonio Zampolli" -CNR –Pisa (Italy) for the invaluable help not only regarding the first profile output of the CLEC corpus but concerning advice and support in the making of our project. Naturally, all errors are our own. We also thank the Teaching Innovation Section of the University of Cádiz for having financed this project.

References

Alanen, R., Huhta, A., and Tarnanen, M. (2010). Designing and assessing L2 writing tasks across CEFR proficiency levels. *Eurosla Monographs Series, 1,* 21 - 56.

Banerjee, J., Franceschina, F., and Smith, A.M. (2004). Documenting features of written language production typical at different IELTS band score levels. *IELTS Research Reports, 7*, Retrieved June 12, 2014 from www.ielts.org.

Barbagli A., Lucisano P., Dell'Orletta F., Montemagni S., and Venturi G. (2014). Tecnologie del linguaggio e monitoraggio dell'evoluzione delle abilità di scrittura nella scuola secondaria di primo grado. *Proceedings of the First Italian Conference on Computational Linguistics (CLiC-it),* 9-10, December, Pisa, Italy.

Biber, D. (1988). *Variation across speech and writing.* Cambridge: Cambridge University Press.

Cimino A., Dell'Orletta F., Venturi G., and Montemagni S. (2013). Linguistic profiling based on general–purpose features and native language identification. *Proceedings of eighth workshop on innovative use of NLP for building educational applications* (pp. 207-215). Atlanta, Georgia, June 13.

Collins-Thompson, K., and Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the american society for information science and technology, 56, 13*, 1448 - 1462.

Council of Europe. (2001). *Common European framework of references for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Dahlmeier, D., Ng, H.T., and Wu, S.M. (2013). Building a large annotated corpus of learner English: The NUS corpus of learner English. *Proceedings of the eighth workshop on innovative use of NLP for building educational applications* (pp. 22-31). Atlanta, Georgia, June13, 2013. Association for computational linguistics.

Dell'Orletta F., Montemagni S., and Venturi G. (2014). 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)*, 163 - 193.

Dell'Orletta, F., and Montemagni, S. (2012). Tecnologie linguistico-computazionali per la valutazione delle competenze linguistiche in ambito scholastico. In S. Ferreri (Ed.), *Linguistica Educativa, Atti del XLIV Congresso Internazionale di Studi della SLI* (pp. 343-359). Roma: Bulzoni Editore.

Dell'Orletta, F., Montemagni, S., and Vecchi, E.M. (2011). Technologie linguistico-computazionali per il monitoraggio della competenza linguistica italiana degli alumni stranieri nella scuola primaria e secondaria. In G.C. Bruno, I. Caruso, M. Sanna, and I. Vellecco (Eds.) *Percorsi Migranti: Uomini, Diritto, Lavoro, Linguaggi* (pp. 319-336). Milano: McGraw-Hill.

Dell'Orletta, F., Montemagni, S., and Venturi, G. (2011). READ-IT: Assessing readability of Italian texts with a view to text simplification. *Proceedings of the workshop on speech and language processing for assistive technologies (SLPAT 2011)* (pp. 73-83). July 30, 2011, Edimburgh, UK.

Dell'Orletta, F., Montemagni, S., and Venturi, G. (2013). Linguistic profiling of texts across textual genres and readability levels. An exploratory study on Italian fictional prose. *Proceedings of recent advances in natural language orocessing* (pp.: 189-197). Hissar, Bulgaria, September 2013.

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

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

Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. *Proceedings of NAACL HLT-2007* (pp. 460-467).

Hendriks, H. (2008). Presenting the English Profile Programme: In search of criterial features. *Research Notes, 33*, 7 - 10.

Hulstijn, J.H., Alderson, J.C., and Schoonen R. (2010). Developmental stages in second-language acquisition and levels of second-language proficiency: Are there links between them? *Eurosla Monographs Series, 1*, 11 - 20.

Kurtes, S., and Saville, N. (2008). The English Profile Programme-An overview. *Research Notes, 33*, 2 - 4.

Montemagni, S. (2013). Technologie linguistico-computazionali e monitoraggio della lingua italiana. *Studi Italiani di Linguistica Teorica Applicata (SILTA), Anno XLII, N.1* (pp. 145-172).

Norris, J. M. (1996). *A validation study of the ACTFL guidelines and the German speaking test*. Unpublished MA dissertation. Honolulu: University of Hawaii.

Norris, J. M., and Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: the case of complexity. *Applied Linguistics, 30, 555 - 578*.

Petersen, S. E., and Ostendorf, M. (2009). A Machine Learning Approach to reading level assessment. *Computer Speech and Language 23,* 89 - 106.

Roark, B., Mitchell, M., and Hollingshead, K. (2007). Syntactic complexity measures for detecting mild cognitive impairment. *Proceedings of ACL workshop on Biological, translational, and clinical language processing (BioNLP'07)* (pp. 1-8). Prague, Czech Republic.

Sagae, K., Lavie, A., and MacWhinney, B. (2005). Automatic measurement of syntactic development in child language. *Proceedings of the annual meeting of the Association for Computational Linguistics (ACL 2005* (pp: 197-204). University of Michigan, USA.

Salamoura, A., and Saville, N. (2010). Exemplifying the CEFR: Criterial features of written learner English from the English Profile Programme. *Eurosla Monographs Series, 1,* 101 - 132.

Van Ek, J. A., and Trim, J. L. M. (1989a). *Threshold*. Council of Europe, Cambridge: Cambridge University Press.

Van Ek, J. A., and Trim, J. L. M. (1989b). *Waystage*. Council of Europe, Cambridge: Cambridge University Press.

Van Ek, J. A., and Trim, J .L. M. (2001). *Vantage*. Council of Europe, Cambridge: Cambridge University Press.

Van Halteren, H. (2004). Linguistic profiling for author recognition and verification. *Proceedings of the Association for Computational Linguistics (ACL04)* (pp. 200-207).

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