JOURNAL OF ENGLISH STUDIES - VOLUME 14 (2016), 215-237. http://doi.org/10.18172/jes.3090

ANALYSING CORPUS-BASED CRITERIAL CONJUNCTIONS FOR AUTOMATIC PROFICIENCY CLASSIFICATION

ÁNGELES ZARCO-TEJADA[1]
CARMEN NOYA GALLARDO[2]
Mª CARMEN MERINO FERRADÁ[3]
ISABEL CALDERÓN LÓPEZ[4]
University of Cádiz
[1]angeles.zarco@uca.es
[2]carmen.noya@uca.es
[3]maricarmen.merino@uca.es
[4]isabel.calderon@uca.es

ABSTRACT. *The linguistic profiling of L2 learning texts can be taken as a model for automatic proficiency assessment of new texts. But proficiency levels are distinguished by many different linguistic features among which the use of cohesive devices can be a criterial element for level distinctions, either in the number of conjunctions used (quantitative) and/or in the type and variety of them (qualitative). We have carried such an analysis with a subgroup of the CLEC (CEFR-levelled English Corpus) using Coh-Metrix, a tool for computing computational cohesion and coherence metrics for written and spoken texts, but our results suggest that automatic proficiency level assessment needs a deeper examination of conjunctions that should rely on the analysis of conjunction-types use and conjunction varieties, with an analysis of lexical choice. A variable based on familiarity ranks could help to predict cohesive levels proficiency-oriented.*

Keywords: Cohesion, language assessment, corpus linguistics, L2 English learning texts, linguistic profiling, Coh-Metrix.

ANÁLISIS BASADO EN CORPUS DE LAS CONJUNCIONES PERTINENTES PARA LA CLASIFICACIÓN AUTOMÁTICA DE LA COMPETENCIA

RESUMEN. *Los estudios de perfil lingüístico de textos de aprendizaje de una segunda lengua pueden ser considerados como modelo para establecer automáticamente evaluaciones de nivel de lengua de textos nuevos. Sin embargo, los niveles de competencia lingüística vienen determinados por múltiples elementos, entre los que el uso de recursos para la cohesión pueden ser considerados como elementos determinantes para establecer diferencias entre niveles, sea por el uso del número de conjunciones (análisis cuantitativo) sea por el del tipo y variedad de ellas (análisis cualitativo). Hemos realizado un análisis con un sub-grupo de textos del CLEC (CEFR-levelled English Corpus) mediante Coh-Metrix, herramienta que computa la cohesión y coherencia de textos escritos y orales, sin embargo los resultados de este análisis sugieren que la evaluación automática de los niveles de competencia necesitaría de un más profundo examen de las conjunciones que tuviera en cuenta los tipos, la variedad y la elección léxica. Así, sugerimos la necesidad de añadir una variable basada en niveles de familiaridad para predecir niveles de cohesión orientados hacia la medida de la competencia lingüística.*

Palabras clave: Cohesión, evaluación de la competencia, lingüística de corpus, textos para el aprendizaje del inglés como L2, perfil lingüístico, Coh-Metrix.

Received 11 October 2016
Revised version accepted 15 November 2016

1. INTRODUCTION

The aim of our study is to analyse which and how cohesive devices discriminate among levels of proficiency and if there are automatic tools that can predict proficiency classifications according to connectors use. For this study we have used the written sub-set of the CLEC corpus (CEFR-levelled English Corpus), a corpus developed at the University of Cádiz for natural language purposes formed by CEFR-levelled learning material texts used in our department. The CLEC is a proficiency-levelled English corpus that covers A1, A2, B1, B2 and C1 CEFR levels and that has been built up to train statistical models for automatic proficiency assessment (Dahlmeier *et al.* 2013; Montemagni 2013; Dell'Orletta *et al.* 2011a, 2011b, 2012 and 2013). We follow an approach that suggests that the identification of discriminating features would be beneficial for establishing boundaries between levels of proficiency and for learning texts proficiency verification and design (Crossley *et al.* 2011; Crossley, Greenfield and McNamara 2008; Graesser *et al.* 2004).

The work we bring up here analyses A2, B1 and B2 levels of the CLEC and it is part of a set of analysis made in order to check and verify if cohesion is appropriately considered in the books used by our learners. As Mahlberg (2006: 107) points out, textbooks have to choose right cohesive categories for learners and she mentions appropriate exemplification, the time-consuming effect of textual analysis, cohesive devices as genre-specific features or generalisation, as essential points that make cohesion a challenging problem within second language texts production. Following Mahlberg (2006) and her warnings on the difficulties to describe cohesion in textbooks, we assume a corpus linguistics approach that can help to analyse texts automatically and outline the most salient cohesive features that distinguish texts. In any case, these tools can help to analyse the actual state of the teaching materials we are using. Thus, in order to achieve our aim we have selected a set of the most representative texts of the written exercises of each level trying to choose the same number of grammatical exercises and short stories with a total number of 10000 words per level. In previous studies we have described how cohesion is achieved in oral and written texts of the CLEC with the AntConc software. Now, following Crossley and McNamara (2009, 2011) and Crosley *et al.* (2009) and their studies on L1-L2 differences regarding lexical cohesion, our goal is to examine written corpora with Coh-Metrix, a system for computing computational cohesion and coherence metrics for written and spoken texts (see section 4.3 for a description of Coh-Metrix), in order to automatically discern among levels of proficiency in terms of their cohesive devices and to identify cohesive devices that are more representative of each level. In this paper we concentrate on conjunctions, one of the four ways to create cohesion according to Halliday's functional grammar principles. We will try to identify if differences among levels of proficiency are based on quantitative or qualitative criteria and will try to establish differences and specify boundaries in terms of textual cohesion. The computational linguistic tools Coh-Metrix 3.0 (McNamara, Louwerse, Cai and Graesser 2013), and AntConc 3.4.3 (Anthony 2014) will be used to analyse our corpora. Establishing differences among levels automatically in terms of textual cohesion using CEFR-levelled learning texts is a first step towards the identification of cohesive devices proficiency oriented and a procedure to verify if and how cohesion is achieved by CEFR-levelled learning materials.

2. STATE OF THE ART

Research on the linguistic profiling of texts has been very fertile over the past decades with many different targets. Many of these are related to linguistic competence and text types assessment: NLP uses for L1 and L2 text readability

measuring (Heilman *et al.* 2007; Collins-Thompson and Callan 2005); authorship identification (McCarthy *et al.* 2006); genre classification or readability levels (Montemagni 2013; Dell'Orletta *et al.* 2013); deficit cognitive analysis through syntax procedures (Roark *et al.* 2007); development of child language through complex syntax use (Sagae *et al.* 2005); text readability measuring with the ranking of documents by reading difficulty or reading abilities as a component of linguistic proficiency (Petersen and Ostendorf 2009); detection of differences between spoken or written English (Louwerse *et al.* 2004). Especially relevant for our study is the role of linguistic features in second language proficiency (Connor 1990; Engber 1995; Ferris 1994 and 2003; Grant and Ginther 2000; Jarvis 2002; McCarthy 2005; Crossley *et al.* 2007) and within this area, the role of cohesion and the use of cohesive devices. Most of the work done on cohesion has focused on the differences between L1 and L2 corpora or on the differences among L2 texts with a learner production analysis goal (Crossley and McNamara 2009; Chen 2008; Granger and Tyson 1996; Green 2012). Among all these, it sets up as a reference for our study Crossley and MacNamara's (2012) analysis on cohesion on L2 writing texts for proficiency matters. Even though their insights deal with L2 analysis and our study analyses proficiency-levelled learning materials, we assume that a similar procedure can be realized.

In fact, one of our aims is to determine the level of cohesion manifested in proficiency-levelled learning texts and compare it to L1 and L2 production statements. There are three main opinions on the relationship proficiency-textual cohesion: high cohesion corresponds to high proficiency (Ferris 1994; Liu and Braine 2005), the relationship cohesion and proficiency is not significant (Johnson 1992; Castro 2004) and high proficiency does not relate to the use of cohesive devices (Crossley and McNamara 2009, 2011, 2012; McNamara *et al.* 2010). According to Crossley and McNamara (2012), Coh-Metrix automatic variables can predict L2 proficiency writing based mainly not on the use of more cohesive devices, but on the use of more linguistically sophisticated terms.

Based on these assumptions and on previous research (Zarco-Tejada *et al.* 2015a), and considering our corpus is formed by learning materials that have a pedagogical function, that is to say, they present a double feature, on the one hand, learning materials try to emulate native English and, on the other, learning materials include proficiency-levelled morphosyntactic, semantic and lexical elements with an academic purpose, we start our research from several hypotheses:

a. Proficiency levels should differ in their cohesive accomplishment: upper proficiency levels should have more-cohesive texts (Collins 1998; De Villez 2003; Witte and Faigley 1981).

b. Proficiency could be related to the number of cohesive devices and to their type (Crossley and McNamara's linguistic sophistication).

c. Automatic tools should show differences among levels in terms of proficiency and in terms of cohesion.

3. LINGUISTIC FRAMEWORK: M.A.K. HALLIDAY

It is an undeniable fact that Halliday and Hasan's *Cohesion in English* is the most relevant and influential study on the notion of cohesion and how it works in English texts since it was published in 1976. As a result, it has been studied, up to now, from different linguistic frameworks, and applied to many other fields such as "stylistics, discourse analysis, language teaching and learning, translation studies, psycholinguistics and sociolinguistics", while still being also very useful to the analysis of texts beyond and around the sentence level as Xi (2010: 141) explains. Moreover, Halliday and Hasan continued developing and improving their theory on cohesion in successive publications.

Given the importance of their studies and conclusions, we have decided to follow their theoretical framework on cohesion in order to analyse the cohesive devices found in the English learning materials we have selected. However, we are going to focus specifically on the conjunction category. To that end, several fundamental concepts in Halliday's work should be first described, though very briefly. The first one is language, which is defined as a system for making meanings by means of wording. According to Halliday and Matthiessen (2014), we are free to make a choice, out of a set of the systems and resources the language has, of the forms that best express what we want to say, in such a way that there is interaction between writer and reader or speaker and addressee. In this sense, language provides a theory of human experience which is transformed into meaning. Halliday identifies three metafunctions or kinds of meaning: the ideational function or language as reflection; the interpersonal function, described as language as action; and the textual function or language as information. The last one is intrinsic to language and deals with the construction of texts, in other words, this is the function the language has to create written or spoken texts that cohere with themselves in the particular situation in which they are used (Halliday and Matthiessen 2014: 30-31). In sum, according to Morley (2000: 13), it organizes the informational content of the clause in a coherent and cohesive way, thus involving the thematic structure, the information structure and cohesion. Systemicists argue that the three metafunctions constitute the functional components of the semantic system that is language, making the three kinds of meanings at the same time.

Eggins points out (2004: 3) that each one expresses a kind of semantic organisation though connected and fused together to produce a single wording.

The second notion, text, described as a unified whole, is the process of making meaning in context, thus referring to any stretch of spoken or written language. Text is not a grammatical unit, but a unit of language in use, so a semantic unit (Halliday and Hasan 1976: 1). Texture, closely related to text, is defined as the property of being a text, that is, any stretch of spoken or written language that holds the clauses together as a unified whole coherently and making sense. It involves the interaction of two groups of resources: structural and cohesive. At the same time, language functions in context, and what the speakers say make sense according to it (Bloor *et al.* 1995: 9). Two types of context are distinguished, context of situation, which regards the immediate social and situational background, and context of culture that refers to an external and broader background described as the contextual potential of a community (Halliday and Matthiessen 2014: 32). Accordingly, we are able to deduce or to predict the context through the meanings and the grammatical choices that have been made. Hence it is inferred, as Thompson states (2004: 9-12), that language and context are interdependent.

We previously indicated that cohesion and coherence constitute crucial notions in our work. Cohesion is described as the interpretation the speaker makes on something else that has previously been mentioned, or is going to be said by reference to another. Consequently, Halliday and Hasan (1976: 4-5) explain that it is a semantic concept and part of the system of a language considering it is realized through the lexicogrammatical system. Therefore, the English language has linguistic resources whose function is to link "an element of language with what has gone before or what follows in a text" (Bae 2001: 55). All grammatical units of any size, sentences, clauses, groups and words can be linked, and so all of them may have cohesive function. Coherence is thus a mental phenomenon that refers to the way they relate to the context. These extra-linguistic elements shaping coherence are divided into three types, namely, field (focusing on the kind and aims of the interaction), mode (referring to the channel of communication) and tenor (focusing on interlocutors and the relationship between them).

There are two types of cohesion: grammatical (reference, conjunction, substitution and ellipsis) and lexical (repetition, synonym, hyponym, collocation, etc.). In this way, five categories of cohesive resources or ties (1976: 3) have traditionally been identified in the English language. However, Halliday and Matthiessen (2014: 52, 612) distinguish within the system of cohesion only four, considering both substitution and ellipsis as just one resource.

As far as conjunction is concerned, it is described in terms of word class as a class within the adverbials, with the function of linking or joining sentences to each other. At the same time it is usually described, as Eggins points out, to be mainly grammatical, though with a lexical component in it (2004: 5). According to Bloor (1995: 98), it is "the term used to describe the cohesive ties between sections of text in such a way as to demonstrate a meaningful relationship between them".

As stated above, we decided to follow Halliday's classification of the system of cohesion (2014: 612-614), based upon the logico-semantic relationships between clauses. We are going to focus specifically on the expansion relationship and the subtypes within it, explained by Downing and Locke (2006: 277) in the following way: "one clause expands another by clarifying or exemplifying (elaboration); by adding or contrasting some feature (extension), or by providing circumstantial information such as time, cause and condition (enhancement)". They allow us to create meaningful structure links between clauses from a semantic point of view.

The first subtype, the semantic relationship of elaboration, is characterised by introducing background information. It is subdivided into apposition and clarification. As a result, the elaborating clause refers to clarifying, specifying, exemplifying, restating, etc. While apposition is composed of appositive, expository and exemplifying conjunctions (*I mean, for example, in other words, for instance, thus,* etc.), clarification comprises clarifying conjunctions, which may be corrective, distractive, dismissive, particularizing, presumptive, summative, or verificative (*at least, by the way, anyway, in particular, actually, in fact, in any case, as a matter of fact, to be more precise, incidentally, in short, briefly, to sum up, etc.).*

The second subtype, extension, refers to extending or contrasting something new, thus providing an exception, or offering an alternative. This is made up of additive conjunctions that can be positive or negative (*and, also, in addition, furthermore, moreover, nor*, etc.); adversative (*but, yet, however, on the other hand, etc.*); and variation with the meaning of replacive, subtractive and alternative (*instead, or, or else, on the contrary, from that, except for that, alternatively*).

Finally, the third subtype, enhancement, in which the enhancing clause presents four circumstancial features: manner, matter, spatio-temporal, and causal-conditional. Manner can be either of comparison or means (*similarly, likewise, in a different way, by such means, etc.*). The second one, matter, can be either positive or negative (*here, there, as to that, in that respect, in other respects, elsewhere*). The third one, called spatio-temporal, refers to time or place (*finally, then, in the end, at once, at that time, apart from that, before that, next time, previously, up to now, lastly, etc.*). Causal-conditional, the last circumstantial type, is broadly subclassified as general (*therefore, so, then, hence, because of that, for*),

and specific (*result, reason, purpose, conditional, positive, negative, concessive*; *as a result, still, though, for that reason, even so, for that purpose, otherwise, in consequence, on account of this, under the circumstances, etc.*).

4. OUR RESEARCH

4.1. MAIN GOALS

Our study is divided in two main analyses. The first one deals with the automatic exploration of our corpus with two main objectives, the distinction of texts according to levels of proficiency based on syntactic, semantic and lexical criteria, and the analysis of cohesion proficiency-determined. The second one deals with identifying qualitative cohesive criteria for proficiency level distinctions.

4.2. DESCRIPTION OF THE CORPUS

We have used the written sub-set of the CLEC corpus (CEFR-levelled English Corpus) (for a description of the corpus see Zarco-Tejada *et al.* 2015b), a corpus developed at the University of Cádiz for natural language purposes formed by CEFR-levelled English learning texts used in our department. The work we bring up here analyses A2, B1 and B2 levels of the CLEC (see Appendix for the list of learning materials used as source) and it is part of a set of analysis made in order to check and verify if cohesion is appropriately considered in the books used by our learners. In order to achieve our aim, we have selected a set of the most representative texts of the written exercises of each level trying to choose the same number of grammatical exercises and short stories with a total number of 10000 words per level.

4.3. COMPUTATIONAL TOOLS

For our first analysis we have used Coh-Metrix 3.0 software (McNamara, Louwerse, Cai, and Graesser 2013), whereas we have used AntConc 3.4.3 (Anthony 2014) for the qualitative study.

4.3.1. Coh-Metrix

This is a system for computing computational cohesion and coherence metrics for written and spoken texts. The variables used in our research have been divided according to the double objectives mentioned above. For a description of many of the features reported by Coh-Metrix see Graesser *et al.* (2004):

- Variables that analyse linguistic complexity:
  - Syntactic simplicity: it reflects the degree to which the sentences in the text contain fewer words and use simpler syntactic structures.
  - Type-Token ratio: TTR (Templin 1957) is the number of unique words (types) divided by the number of tokens of these words. When the value approaches to 1, each word occurs only once. As the Type/Token ratio decreases, words are repeated many times which increases the ease of text processing.
  - Familiarity: it rates how familiar a word is for an adult. Sentences with more familiar words are processed more quickly.
  - Hypernymy: a lower value reflects the use of less specific words, while a higher value reflects the use of more specific words.
  - Readability: it assesses texts on difficulty.
  - Occurrence of words before the main verb: it rates the mean number of words before the main verb of the main clause.
  - Occurrence of modifiers per NP: it rates the mean number of modifiers per NP.
  - Syntactic similarity of adjacent sentences: it rates the proportion of intersection tree nodes between all adjacent sentences and across paragraphs.
- Variables that analyse cohesion:
  - Deep cohesion: this variable measures the number of causal and intentional connectives. The more number of connectives the better coherence and understanding of causal events and processes in the text.
  - Connectivity: it reflects the degree to which the text contains explicit adversative and comparative connectives to express relations in the text. It reflects the number of logical relations.
  - All connectives: this variable measures the incidence of all connectives. Connectives are a way to create cohesive links within the text (Cain and Nash 2011; Crismore, Makkanen and Steffensen 1993; Longo 1994; Sanders and Noordman 2000; van de Kopple 1985).
  - Causal connectives: this variable measures the incidence of causal connectives such as *so that, because, since*, etc.
  - Adversative/contrastive connectives: this variable measures the incidence of adversative/contrastive connectives such as *but, however, yet, still*, etc.
  - Temporal connectives: this variable measures the incidence of temporal connectives such as *in the end, next, then, now*, etc.
  - Additive connectives: this variable measures the incidence of additive connectives such as *and, moreover, furthermore, therefore*, etc.

4.3.2. AntConc

This software has been used to analyse qualitatively most of the conjunctions that contribute to create cohesion. We have analysed them individually regarding frequency and concordances for each level. With this approach we check our hypothesis that sets out that cohesion proficiency-oriented could be related to lexical choice.

5. QUANTITATIVE ANALYSIS

We have grouped the output in several tables as variables applied relate to linguistic complexity or to cohesion specifically. As regards linguistic complexity, nine variables have been selected: syntactic simplicity, lexical diversity, familiarity of content words, hypernymy of nouns and verbs, reading ease, words before main verbs, modifiers per noun, sentence syntax similarity between adjacent sentences and sentence syntax similarity across paragraphs. With this first analysis we wanted to have a first Coh-Metrix analysis of our sub-corpus of CLEC and check if levels of proficiency are reflected in linguistic feature choices that could be detected by Coh-Metrix variables.

Table 1. Linguistic complexity variable measuring: A2, B1 and B2 written sub-corpus of CLEC with Coh-Metrix.

CEFR levels	**A2**	**B1**	**B2**
Linguistic complexity			
Syntactic simplicity	0.80	0.664	0.063
Lexical diversity (Type/Token ratio)	0.74	0.815	0.867
Lexical diversity (MTLD all words)	51.78	74.62	71.72
Familiarity of content words	574.158	583.74	585.811
Hypernymy of nouns and verbs	1.602	1.494	1.395
Reading ease	75.260	78.936	80.321
Words before main verbs	1.590	2.068	1.973
Modifiers per noun	0.561	0.580	0.618
Sentence syntax similarity between adjacent sentences	0.199	0.150	0.124
Sentence syntax similarity across paragraphs	0.199	0.146	0.119

Results are according to predictions. The table above shows how B2 texts are syntactically more complex as variables show an upward tendency from A2 to B2, or a downward tendency in the case of the syntactic simplicity variable. As the

Type/Token ratio indicates, lexical diversity is higher in upper levels. As the value approaches to 1, each word is used only once and thus lexical diversity is higher, which implies a lower cohesion of the text. This value is related to linguistic proficiency even though it has a reading on the cohesion perspective: the higher the Type/Token value the lower cohesion is achieved. Such result is supported by values obtained for "content word overlap in adjacent sentences" (A2: 0.128; B1: 0.121; B2: 0.113) and for "content word overlap in all sentences" (A2: 0.117; B1: 0.105; B2: 0.096) with a decreasing tendency from A2 to B2. According to these values, texts from lower levels repeat words more often than upper proficiency level texts, a fact that can be related to less lexical diversity in low levels but achievement of cohesive texts through repetition of tokens of the same type. Diversity of words is analysed by the Measure of Textual and Lexical Diversity (MTLD) variable too. In this case, the difference is evident between A2 and B1 levels but the output for B2 is lower (A2: 51.78; B1: 74.62; B2: 71.72). As far as we are concerned, lexical diversity is a variable that can be interpreted in two ways as regards textual cohesion. On the one hand, the use of different terms implies less cohesive texts, a feature of low proficiency levels, but, on the other, lexical diversity is produced by the use of a wider vocabulary which is a feature of higher proficiency texts, in the same line as Crossley and McNamara's (2011) linguistic sophistication concept.

The output of variables such as hypernymy, reading ease, words before main verbs or modifiers per noun and sentence syntax similarity, show how complexity is a proficiency feature with higher values in upper levels. In the case of the hypernymy variable, the scores (A2: 1.602; B1: 1.494; B2: 1.395) indicate the use of less specific words in upper levels and thus, more difficult to be processed. The number of words before the main verb or within the NP shows the use of complex phrase structures in upper levels. Finally, another feature of linguistic complexity is determined by sentence syntax similarity (A2: 0.199; B1: 0.150; B2: 0.124), that analyses syntactic uniformity. Upper levels show lower uniformity of syntactic constructions and thus more complex syntax.

In order to account for the linguistic cohesion of texts, we have selected seven Coh-Metrix variables. Deep cohesion analyses the degree to which the text contains causal and intentional connectives and the variable connectivity measures the explicit adversative, additive and comparative connectives in the text. Besides, we have analysed the more specific connectives e.g. causal, adversative, temporal and additive, individually. Results are listed in table 2 below.

Table 2. Linguistic cohesion measuring: A2, B1 and B2 written sub-corpus of CLEC with Coh-Metrix.

CEFR levels	A2	B1	B2
Linguistic cohesion			
Deep cohesion	-0.33	0.129	0.343
Connectivity	-1.45	-1.419	-0.90
Connectives			
All connectives	63.213	67.50	84.18
Casual connectives	20.318	19.169	29.074
Adversative and contrastive connectives	11.442	15.458	12.434
Temporal connectives	12.993	13.469	18.434
Additive connectives	31.009	35.034	27.462

The first two variable outputs, which are general, indicate that upper proficiency levels show a gradual higher level of cohesion in terms of explicit connectives use, figures that are supported by the general variable "all connectives" (A2: 63.213; B1: 67.50; B2: 84.18). Regarding connective categories, the upward tendency of connective use of higher proficiency levels is reflected by casual and temporal connectives whereas adversative and additive connectives results do not show a unified tendency.

The main question now is how these results can be related to cohesive procedures of upper proficiency texts. In other words, can we support proficiency on quantitative analysis of cohesive devices only? Scores on additive connectives show higher results in lower levels of proficiency (A2 and B1 than B2), whereas the output of adversative connectives is higher for B1 than for B2 (B1: 15.458; B2: 12.434).

Our hypothesis b, that follows Crossley and MacNamara's linguistic sophistication concept, leads our research towards a more specific analysis of cohesive devices, a qualitative one, in order to be able to establish cohesive differences among levels. From an intuitive approach, we can imagine texts that show less scores in terms of the total number of connectives used but that belong to upper levels because of the use of a wider and more varied set of terms. It is for this reason that we start a qualitative analysis in the following section including in the table connectives with scores ≥ 1 and, thus, leaving aside all those that, having been searched for, do not show any result.

6. QUALITATIVE ANALYSIS

We include below three tables according to Halliday's (2004) taxonomy, with the three main types of conjunctions: elaborating, extending and enhancing conjunctions. Within them, appositive, clarifying, additive, adversative, varying, matter, manner, spatio-temporal and causal-conditional categories have been considered. The tables display the number of hits found in the sub-set of texts under analysis with the AntConc system. The concordance layer gave us the number of hits found and we used the context to eliminate ambiguous examples manually (i.e. "so" as connector or intensifier "it's so good to see you"). The tables will be commented separately for a better explanation of facts and a more general consideration will be made at the end.

6.1. ELABORATING CONJUNCTIONS

Table 3. Number of hits of the conjunctions analysed in A2, B1 and B2 levels of written English of CLEC: elaborating conjunctions.

CEFR levels	A2	B1	B2
Appositive conjunctions			
I mean	0	0	1
Thus	0	0	1
For example	0	1	1
Clarifying conjunctions			
at least	0	0	2
anyway	0	2	3
actually	1	2	0
In fact	1	0	1

According to data, the learning texts under analysis use a very low number of conjunctions of the same category as well as the number of individual conjunctions is very small. Considering the eight appositive conjunctions studied ("in other words", "that is", "I mean", "to put it another way", "thus", "for example", "for instance", "to illustrate"), A2 texts show 0%, B1: 0.125% and B2: 0.375% of individual conjunctions use. Similar results can be found regarding clarifying conjunctions. With regard to the 21 conjunctions under analysis ("or rather", "at least", "to be more precise", "by the way", "incidentally", "in any case", "anyway", "leaving that aside", "in particular", "more especially", "to resume", "as I was saying", "to get back to the point", "in conclusion", "in short", "briefly", "to sum up", "actually", "verificative", "as

a matter of fact", "in fact"), A2 shows an average rate of 0.095%, B1: 0.095% and B2: 0.142% of individual conjunctions realizations. Even though the number of hits found is very small and the variety use too, there is an upward tendency from low proficiency levels to upper levels. In fact, B2 shows more number of hits and more conjunction variety use than B1 and B1 than A2.

6.2. EXTENDING CONJUNCTIONS

Table 4. Number of hits of the conjunctions analysed in A2, B1 and B2 levels of written English of CLEC: extending conjunctions.

CEFR levels	A2	B1	B2
Additive conjunctions			
And	245	273	265
Also	8	4	5
Nor	0	0	3
Adversative conjunctions			
But	55	87	71
Yet	4	8	3
However	1	3	3
Varying conjunctions			
Instead	0	2	1
Apart from that	0	0	1
Or	21	18	8

Results obtained for extending conjunctions are different from the previous one. There are three conjunctions that are very much used in the three levels of proficiency, "and", "but" and "or". Having a look at each category specifically, A2 has an average score of 0.333%, B1: 0.333% and B2: 0.5% regarding additive conjunctions, having considered 6 conjunctions ("in addition", "and", "also", "moreover", "furthermore", and "nor"). Within the adversative category, the average 0.75% is the same for the three levels since three conjunctions out of four analysed ("but", "yet", "however", and "on the other hand") have been found. Finally, the output in varying conjunctions is lower than in the previous two. The average rate is A2: 0.125%, B1: 0.25% and B2: 0.375%, having analysed 8 conjunctions ("on the contrary", "instead", "on the other hand", "apart from that", "except for that", "or", "or else", and "alternatively").

According to results, learning books show a higher number of extending conjunctions than elaborating conjunctions in texts, and, especially, three

conjunctions mentioned above are very much used. Besides this, results show that there is an upward tendency from low levels to upper levels of proficiency in terms of the overall conjunctions use as well as on the conjunction use variety. Upper levels show a wider range of conjunction examples.

6.3. ENHANCING CONJUNCTIONS

Table 5. Number of hits of the conjunctions analysed in A2, B1 and B2 levels of written English of CLEC: enhancing conjunctions.

CEFR levels	A2	B1	B2
Matter conjunctions			
Here	34	8	29
There	38	11	21
Elsewhere	0	0	1
Manner conjunctions			
Spatio-temporal conjunctions			
Afterwards	0	0	1
Then	1	17	19
Next	1	21	7
First	3	8	14
Just	3	17	32
Now	5	13	23
Finally	0	2	9
In the end	1	2	0
So	16	22	27
Next time	2	0	0
At that time	0	1	0

CEFR levels	A2	B1	B2
Causal-conditional conjunctions			
So			
Then	17	19	26
As a result	0	0	1
Then			
Otherwise	0	0	1
If not	0	0	1
Yet	0	8	3
Still	0	5	11
Though	0	3	4
However	1	3	3

Enhancing conjunctions are analysed according to 4 categories: matter, manner, spatio-temporal and causal-conditional. Results differ from one category to another. On the one hand, the most salient result is the one regarding manner conjunctions with no hits at all. Conjunctions such as "likewise", "similarly", "in a different way", "by such means", "in the same manner" or "thereby" are not found in A2, B1 or B2 learning materials under study. On the other, the rest of conjunctions are found in texts with different outputs. The average use of the matter class reaches the average 50% of the 6 conjunctions analysed ("here", "there", "as to that", "in that respect", "in other respect" and "elsewhere"), with 0.333 % for A2 and B1 and 0.5 % for B2. Spatio-temporal conjunctions are very much used not only in terms of the overall number of hits, but in terms of variety. Thus, having analysed 29 conjunctions ("afterwards", "then", "next", "first", "at the same time", "just", "now", "previously", "up to now", "before that", "finally", "lastly", "in the end", "straightaway", "at once", "thereupon", "so", "after a while", "on another occasion", "next time", "an hour later", "next day", "that morning", "all that time", "meanwhile", "at that time", "until then", "up to that point", and "at this moment"), the three levels of proficiency show the same average rate: 0.310% (9 types out of 29) with small differences regarding the total number of hits per level in each type and the variety use. Finally, the causal-conditional group shows an upward tendency in the total number of hits and in the variety of conjunctions. A2 shows an average rate of 0.173%, B1: 0.304% and B2: 0.434%, considering 23 conjunctions ("because of that", "so", "then", "therefore", "hence", "in consequence", as a result", "in account of this", "for that reason", "for that purpose", "under the circumstances", "then", "in that case", "otherwise", "if not", "yet", "still", "though", "nevertheless", "despite this", "however", "even so", and "all the same").

7. ON THE COHESION-PROFICIENCY RELATION

After having analysed our corpus and having produced quantitative and qualitative data, several conclusions can be drawn:

1. Automatic analyses in terms of general scores, as the Coh-Metrix system produces, give very interesting insights on the cohesive status of a text. Deep cohesion, lexical diversity, connectivity and all connectives, in general and individually, are variables that can score texts according to cohesive devices use. The point for us, though, is not the question of which text is more cohesive, but which text belongs to which proficiency level regarding cohesion. We cannot forget we are dealing with texts used for English learning activities and thus texts are presumed to be cohesive.

2. Conjunctions are cohesive devices that help to determine the rank of cohesion achieved in a text. According to our analysis, the upward tendency of connective use of higher proficiency levels is reflected by casual and temporal connectives, whereas adversative and additive connectives results do not show a unified tendency. The question now could be, is a text more or less cohesive because it has more or less explicit connectives, or because it has more or less connectives of different types, or because it has more or less connectives of different categories?
3. A text can be cohesive but the grade of cohesion does not necessarily have to be related to the grade of proficiency. If we have a look at the results displayed by the qualitative analysis, the main differences among levels have to be with the conjunctions types use (lower levels show less conjunctions in terms of varieties), with the conjunctions categories (no examples for manner conjunctions and very low examples for clarifying or varying conjunctions), and with the use of more particular conjunctions (linguistic sophistication). Elaborating conjunctions are used to re-present or make more precise some elements in the discourse. This linguistic ability is found more in upper levels than in lower ones. Extending conjunctions are highly represented by the conjunctions "and", "but" and "or" with high rates, though variety use is very low. Finally, differences among levels regarding enhancing conjunctions are sustained by the low score of manner and matter conjunctions and by the upward tendency in conjunction variety use of spatio-temporal and causal-conditional conjunctions for upper levels of proficiency.
4. Our analysis shows how proficiency level differences in terms of conjunction use are related to two main variables: conjunction variety and conjunctions examples. Results indicate that general quantitative outcomes do not explain proficiency classifications sufficiently. A familiarity variable, as the one used by Coh-Metrix for the whole vocabulary of each text, that could analyse conjunctions only, seems to be a desirable variable to account for cohesion with proficiency classification purposes. A variable that could analyse conjunctions and measure how 'sophisticated' they are, based on familiarity ranks, could help to predict cohesive levels proficiency-oriented.

8. CONCLUSIONS

In this paper we have shown how proficiency levels are distinguished in their use of cohesive devices and how this analysis can be done automatically by Coh-Metrix. The study posed here describes a quantitative and qualitative analysis of

a CEFR-levelled written sub-corpus of the CLEC, with variables that apply to linguistic complexity and to cohesion. Results show that automatic proficiency level distinctions based on the automatic analysis of cohesion would need a deeper examination of conjunctions that could rely on the analysis of conjunction-types use and conjunction varieties, with an analysis of lexical choice. Our analysis has shown that variety and type play an important role in proficiency level distinctions and that such vocabulary differences are related to high linguistic competence. In this sense, cohesive devices cannot be evaluated in quantitative terms only but qualitative criteria can be criterial for level classifications. In line with Crossley and McNamara's linguistic sophistication concept, we suggest that a new variable on the familiarity rank of conjunctions could help to establish proficiency differences determined by cohesive devices use.

Our future research within the analysis of second language texts will focus on the relationship between discourse, cohesion and communication skills under the CEFR specifications for the learner's competence development.

REFERENCES

Anthony, L. 2014. AntConc 3.4.3 [Computer Software]. Tokyo, Japan: Waseda University. <http://www.laurenceanthony.net> (Accessed 6 September 2015).

Bae, J. 2001. "Cohesion and coherence in children's written English: immersion and English-only classes". *Issues in Applied Linguistics* 12 (1): 55-88. <http://escholarship.org/uc/item/3hb2z0s4> (Accessed 15 January 2016).

Bloor, Th. and M. Bloor. 1995. *The Functional Analysis of English: a Hallidayan Approach.* London: Arnold.

Cain, K. and H. M. Nash. 2011. "The influence of connectives on young readers' processing and comprehension of text". *Journal of Educational Psychology* 103 (2): 429-441.

Castro, C. D. 2004. "Cohesion and the social construction of meaning in the essays of filipino college students writing in L2 English". *Asia Pacific Education Review* 5 (2): 215-225.

Chen, J. 2008. "An investigation of EFL students' use of cohesive devices". *Asia Pacific Education Review* 5 (2): 215-225.

Collins, J. L. 1998. *Strategies for Struggling Writers*. New York: Guilford.

Collins-Thompson, K. and J. Callan. 2005. "Predicting reading difficulty with statistical language models". *Journal of the American Society for Information Science and Technology* 56 (13): 1448-1462.

Connor, U. 1990. "Linguistic/rethorical measures for international persuasive student writing". *Research in the Teaching of English* 24: 67-87.

Crismore, A., Markkanen, R. and M. Steffensen. 1993. "Metadiscourse in persuasive writing: a study of texts written by American and Finnish university students". *Written Communication* 10: 39-71.

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

Crossley, S. A., Louwerse, M., McCarthy, P. M. and D. McNamara. 2007. "A linguistic analysis of simplified and authentic texts". *Modern Language Journal* 91: 15-30.

Crossley, S. A. and D. McNamara. 2009. "Computational assessment of lexical differences in L1 and L2 writing". *Journal of Second Language Writing* 18: 119-130.

Crossley, S. A. and D. McNamara. 2011. "Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication". *Journal of Research in Reading* 35 (2): 115-135.

Crossley, S. A., Salsbury, T. and D. McNamara. 2009. "Measuring second language lexical growth using hypernymic relationships". *Language Learning* 60 (3): 307-334.

Crossley, S. A., Salsbury, T., McNamara, D. S. and S. Jarvis. 2011. "What is lexical proficiency? Some answers from computational models of speech data". *TESOL Quarterly* 45 (1): 182-193.

Dahlmeier, D., Ng, H. T. and S. M. Wu. 2013. "Building a large annotated corpus or learner English: The NUS corpus of learner English". *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta, Georgia, June 13, 2013. Association for Computational Linguistics. 22-31.

De Villez, R. 2003. *Writing: Step by Step*. Dubuque, I. A.: Kendall Hunt.

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

Dell'Orletta, F., Montemagni, S. and G. Venturi. 2011b. "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)*. July 30, 2011, Edimburgh, UK. 73-83.

Dell'Orletta, F. and S. Montemagni. 2012. "Tecnologie linguistico-computazionali per la valutazione delle competenze linguistiche in ambito scholastico". *Linguistica*

Educativa. Ed. S. Ferreri. Atti del XLIV Congresso Internazionale di Studi della SLI, Roma, Bulzoni Editore. 343-359.

Dell'Orletta, F., Montemagni, S. and G. Venturi. 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 Processing*. September 2013, Hissar, Bulgaria. Association for Computational Linguistics. 189-197.

Downing, A. and P. Locke. 2006. *A University Course in English Grammar*. London: Routledge.

Eggins, S. 2004. *An Introduction to Systemic Functional Linguistics*. London: Continuum.

Engber, C. A. 1995. "The relationship of lexical proficiency to the quality of ESL compositions". *Journal of Second Language Writing* 4 (2): 139-155.

Ferris, D. 1994. "Lexical and syntactic features in ESL writing by students at different levels of L2 proficiency". *TESOL Quarterly* 28 (2): 414-420.

Ferris, D. 2003. *Response to Student Writing: Implications for Second Language Students.* Mahwah, N.J.: Lawrence Erlbaum.

Graesser, A., McNamara, D. S., Louwerse, M. and Z. Cai. 2004. "Coh-Metrix: Analysis of text on cohesion and language". *Behavioral Research Methods, Instruments, and Computers* 36: 193-202.

Granger, S. and S. Tyson. 1996. "Connector usage in the English essay writing of native and non-native EFL speakers of English". *World Englishes* 15 (1): 17-21.

Grant, L. and A. Ginther. 2000. "Using computer-tagged linguistic features to describe L2 writing differences". *Journal of Second Language Writing* 9: 123-145.

Green, C. 2012. "A computational investigation of cohesion and lexical network density in L2 writing". *English Language Teaching* 5 (8): 57-69.

Jarvis, S. 2002. "Short texts, best fitting curves and new measures of lexical diversity". *Language Testing* 19: 57-84.

Johnson, P. 1992. "Cohesion and coherence in compositions in Malay and English". *Journal of Language Teaching and Research* 23 (2): 1-17.

Halliday M. A. K. 2004. *An Introduction to Functional Grammar*. London: Arnold.

Halliday, M. A. K. 2013. *Halliday's Introduction to Functional Grammar* (4th ed). London: Routledge.

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

Halliday, M. A. K. and C. Matthiessen. 2014. *Halliday's Introduction to Functional Grammar*. London: Edward Arnold.

Heilman, M., Collins-Thompson, K., Callan, J. and M. Eskenazi. 2007. "Combining lexical and grammatical features to improve readability measures for first and second language texts". *Proceedings of NAACL Human Language Technologies-2007*. Rochester, New York, 2007. Association for Computational Linguistics. 460-467.

Liu, M. and G. Braine. 2005. "Cohesive features in argumentative writing produced by Chinese undergraduates". *System* 33: 623-636.

Longo, B. 1994. "The role of metadiscourse in persuasion". *Technical Communication* 41: 348-352.

Louwerse, M. M., McCarthy, P. M., McNamara, D. S. and A. C. Graesser. 2004. "Variation in language and cohesion across written and spoken registers". *Proceedings of the 26th Annual Cognitive Science Society*. Eds. K. Forbus, D. Gentner and T. Regier. Mahwah, HJ: Erlbaum. 843-848.

Mahlberg, M. 2006. "Lexical Cohesion. Corpus Linguistic Theory and its Applications in ELT". *Special issue of the International Journal of Corpus Linguistics* 11 (3): 363-383.

McCarthy, P. M. 2005. "An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)". *Dissertation Abstracts International,* 66 (12), UMI No. 3199485.

McCarthy, P. M., Lewis, G. A., Dufty, D. F. and D. S. McNamara. 2006. "Analyzing writing styles with Coh-metrix". *Proceedings of the Florida Artificial Intelligence Research Society International Conference*. Eds. G. Sutcliffe and R. Goebel. AAAI Press. 764-769.

McNamara, D. S., Louwerse, M. M., Cai, Z. and A. Graesser. 2005. Coh-Metrix version 1.4. <http//:cohmetrix.memphis.edu> (Accessed September 2015).

McNamara, D. S., Crossley, S. A. and P. McCarthy. 2010. "Linguistic features of writing quality". *Written Communication* 27 (1): 57-86.

McNamara, D. S., Louwerse, M. M., Cai, Z. and A. Graesser. 2013. Coh-Metrix version 3.0. <http://cohmetrix.com > (Accessed September 2015).

Montemagni, S. 2013. "Tecnologie linguistico-computazionale e monitoraggio della lingua italiana". *Studi Italiani di Linguistica Teorica e Applicata (SILTA)* 42 (1): 145-172.

Morley, G. D. 2000. *Syntax in Functional Grammar. An Introduction to Lexicogrammar in Systemic Linguistics*. NewYork: Continuum. <http://staff.uny.ac.id/sites/default/files/pendidikan/Lusi%20Nurhayati,%20S.Pd.,%20M.Appl.Ling%20(TESOL)/SYNTAX%20IN%20FUNCTIONAL%20GRAMMAR.pdf> (Accessed February 2016).

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

Roark, B., Mitchell, M. and K. Hollingshead. 2007. "Syntactic complexity measures for detecting mild cognitive impairment". *Proceedings of ACL Workshop on Biological, Translational, and Clinical Language Processing (BioNLP'07)*. Prague, Czech Republic. 1-8.

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

Sanders, T. J. M. and L. G. M. Noordman. 2000. "The role of coherence relations and their linguistic markers in text processing". *Discourse Processes* 29 (1): 37-60.

Templin, M. 1957. *Certain Language Skills in Children*. Minneapolis, MN: University of Minnesota Press.

Thompson, G. 2004. *Introducing Functional Grammar*. New York: Saint Martin's Press.

Van de Kopple, W. J. 1985. "Some explanatory discourse on metadiscourse". *College Composition and Communication* 36: 82-93.

Witte, S. P. and L. Faigle. 1981. "Coherence, cohesion and writing quality". *College Composition and Communication* 22: 189-204.

Xi, Y. 2010. "Cohesion studies in the past 30 years: development, application and chaos". *The International Journal-Language Society and Culture* 31. <www.educ.utas.edu.au/users/tie/JOURNAL> (Accessed January 2016).

Zarco-Tejada, M. A., Noya Gallardo, C., Merino Ferradá, M. C. and I. Calderón López. 2015a. "2L English texts and cohesion in upper CEFR levels: a corpus-based approach". *Procedia – Social and Behavioral Sciences*, 212: 192-197. http://doi.org/10.1016/j.sbspro.2015.11.319.

Zarco-Tejada, M. A., Noya Gallardo, C., Merino Ferradá, M. C. and I. Calderón López. 2015b. "Building a corpus of 2L English for automatic assessment: The CLEC corpus". *Procedia -Social and Behavioral Sciences* 198: 515-525. http://doi.org/10.1016/j.sbspro.2015.07.474.

APPENDIX

List of learning materials used in the present analysis:

A2: A2 New Headway Elementary Student's Book
A2 ACTIVATE Workbook with key
Total number of words: 10052

B1: B1 New Headway Intermediate Student's Book
B1-B2 Grammar in Use
B1 New Headway Pre-Intermediate Student's Book
Total number of words: 10036

B2: TELC Mock ExaminationB2
B2+ Grammar Practice for Upper Intermediate Students
B2 New English File Upper Intermediate Student's Book
Total number of words: 10228