Building a Corpus of 2L English for 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, Spain

Financed by the Teaching Innovation Section of the University of Cádiz &
in collaboration with the Istituto di Linguistica Computazionale-(CNR)-Pisa

The CLEC corpus

- What is it? corpus of English as a 2L classified according to CEFR proficiency levels.
- What for? to train statistical models for automatic proficiency assessment.
- How? As a teaching innovation technique.
- Why?

The needs to build up CLEC

- Social demand: grant our students with a language proficiency certificate.
- Problem in our University: production of 2L English materials for language proficiency assessment.
- Solution: automatic classification of texts.

Theoretical backgroud

- Our 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 et al. (2013).
- Dahlmeier et al. (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.

Our project

- Our project was set up in 2012.
- CLEC (CEFR-Labeled English Corpus) .
- More than 200.000 words of grammatical English examples
- Source: 2L English texts already classified for the CEFR levels A1, A2, B1, B2, C1 and C2.
- Texts have been manually encoded .
- The corpus has been annotated with additional information as metadata.

Two goals in *the making* of CLEC

- The production of 2L English materials for language proficiency assessment.
- The application of teaching innovation actions.

NLP Procedure

- 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.
- 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.

Description of CLEC

- *People involved:*
  - *4 teachers*
  - *1 PhD student*
  - *1 Post-graduate student*
  - *10 Collaborating students of the English Studies degree.*
  - *30 undergraduate students of English Studies*

Data collection

- Year 2012-13: 60723 words distributed in the following CEFR levels:
  - A1: 3744 words
  - A2: 20322 words
  - B1: 35383 words
  - B2: 1274 words

- Year 2013-14: 105949 words
  - A1: 3744 words
  - A2: 21239 words
  - B1 45864 words
  - B2: 11189 words
  - C1: 3648 words
  - C2: 20265 words

- Year: 2014-2015: 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.
  - A1: 3744 words
  - A2: 21239 words
  - B1 79923 words
  - B2: 48088 words
  - C1: 64699 words
  - C2: 20265 words

Codifying process

- Optional activity of the Teaching Innovation Program held at the University of Cádiz.
- Students were aware of being part of a research process: improve their University academic training.
- This project was born in collaboration with ILC-Pisa: it was more attractive for students.

Codifying process: 2L English sources

- The student's book sets (from pre-intermediate to advanced) of:
  - New Headway
  - New English File
  - Face2face

Codifying process: text codification procedure

- 1st phase: Exercises done as homework activities.
- 2nd phase: results are checked for grammatical errors.
- Each text is saved in a different file.
- Each text is saved as a plain text.
- Each text is uploaded to the e-learning platform “Corpus”.

Text structure

- Two tabs:
  - Opening tab:
    - Id (identification): source teaching text information
    - Cat (category): linguistic function
    - Arg (argument): grammatical information
  - Closure tab

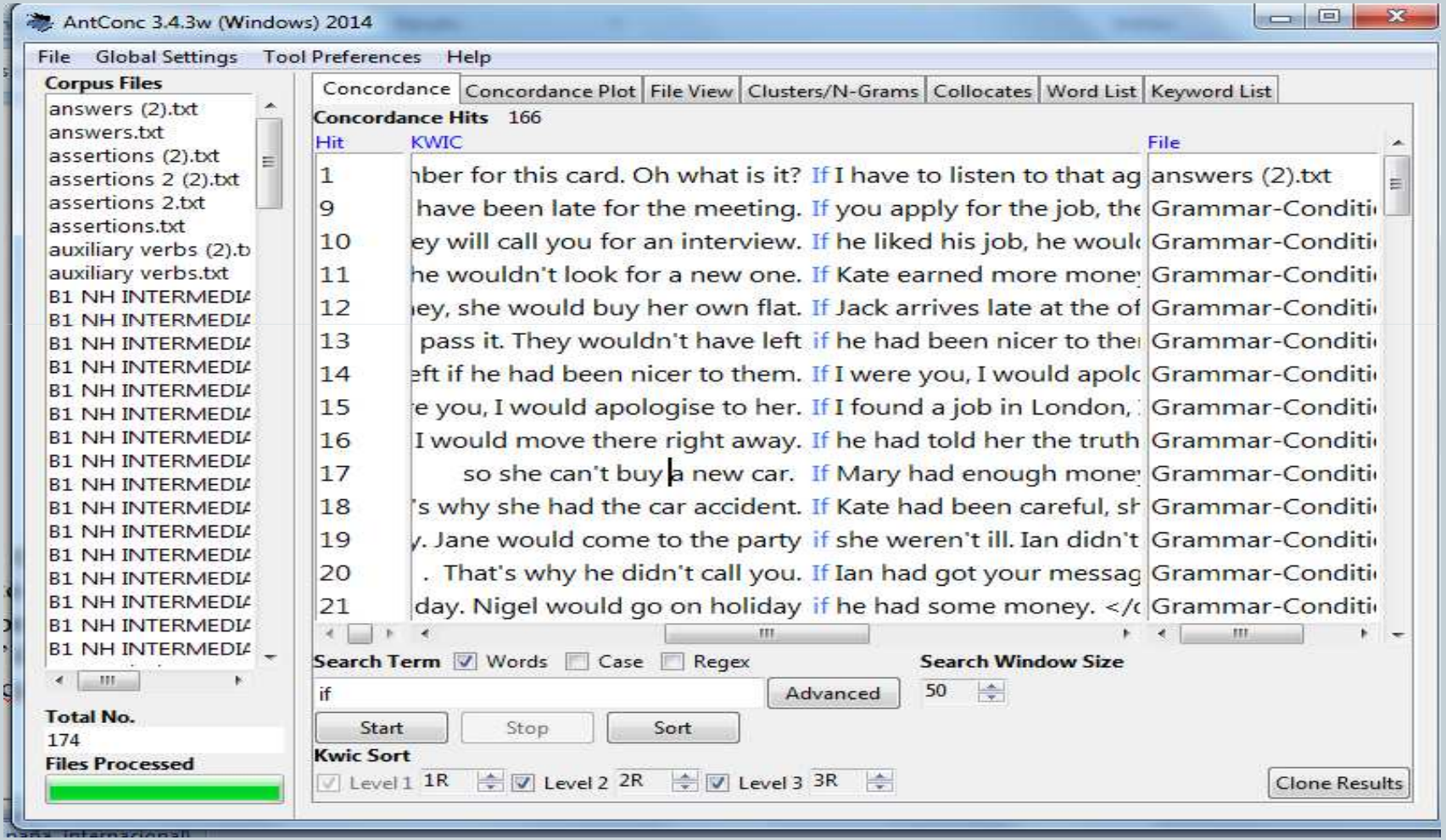
Text tabs examples

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

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

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

KWIC conditional sentences encoded in B1 level of the CLEC and processed by AntConc 3.4.3



Main obstacles in the CLEC construction

- Unbalanced collection of texts.
- Supervising the student's assignments: a hard-working activity.
- Linguistic function identification: inconsistencies.
- Oral English examples.

Linguistic profiling of CLEC: first results

- First step towards our ultimate aim of producing automatic proficiency assessment of new texts.
- The linguistic analysis of our corpus can help to identify and define the criterial CEFR levels features.
- The first outcome is applied to levels A2, B1 and B2 of written English and to B1 and B2 levels of oral English.

Similar approaches

- The linguistic profiling of these levels follows the methodology and linguistic description explained in Montemagni (2013): 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.
- The occurrences of the selected linguistic features are counted for the identification of the text profile (Biber, 1988; van Halteren, 2004).

Linguistic profiling

- 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.

Linguistic profiling

- Differences among levels of proficiency are based on text readability complexity.
- Either lexical or syntactic complexity are analyzed.

Readability measures

- Token-dependent distance is one aspect of readability measure (Lin, 1996; Gibson, 1998).
- Syntactic tree depth is considered a central aspect for text readability assessment (Yngve (1960), Frazier (1985) and Gibson (1998)) .
- Syntactic 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.

Readability measures

- lexical complexity is another criteria that determines readability measures and that we have formalized in terms of:
  - the number of tokens each sentence has.
  - the number of characters within tokens.
  - the type/token ratio that reflects lexical variation in a corpus.

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

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

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

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

Results

- 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.
- Syntactic area:
  - 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.

Conclusions

- As expected, either the lexical features or the syntactic ones show deeper levels of complexity in higher levels of proficiency.
- 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.