Le Corpus d'Étude des Langues Vivantes Appliquées à une Spécialité (CELVA.Sp) source de visualisations linguistiques

Thomas Gaillat, Pascale Janvier, Bénédicte Dumont, Antoine Lafontaine, Anas Knefati

CELVA.Sp : A corpus for the visualisation of linguistic profiles in language learners ⋆

Thomas Gaillat[1], Antoine Lafontaine[2], Anas Knefati[3], Bénédite Dumont[2], Pascale Janvier[2], Rana Challah[1], and Claude Hamon[1]

[1] University of Rennes 2
[2] University of Rennes 1
[3] ENSAI

Résumé This paper presents the design of a Language for Specific Purposes(LSP) Corpus and its exploitation as a source for real-time visualisation of linguistic complexity in learner writings. The corpus is provided with a Natural Language Processing (NLP) tool, called VizLing, used to compute and visualise complexity metrics. The resulting data set is made up of learner writings, metadata and complexity metrics.

Keywords: Learner corpus · Language for Specific Purposes · Automatic analysis · linguistic complexity

1 Introduction

In this chapter, we present the design of a Language for Specific Purposes (LSP) Corpus and its exploitation as a source for real-time visualisation of linguistic complexity in learner writings. We show how a specifically designed corpus can be exploited with Natural Language Processing (NLP) tools to compute and visualise complexity metrics for learner writings.

Learner corpora have been mainly used for two purposes. First they have helped syllabus designers in determining linguistic objectives to focus on. Second, They have been used in class by students by way of corpus exploration activities. However learner corpora have not been exploited much in iCALL systems, except in the area of error detection. Our proposal is to exploit a corpus of English for Specific Purposes (ESP) as a gold standard for the automatic analysis of learner writings within an iCALL system.

Our approach is to build an ESP corpus, called CELVA.Sp [4], in order to subsequently integrate it in a system dedicated to linguistic feedback about the positive linguistic properties of learner language. The assumption is that positive properties vary according to proficiency levels (Hawkins and Filipović, 2012). To achieve this, we have computed linguistic complexity metrics for each of the collected writings of the English subset of the corpus.

⋆. Supported by the DUNE DESIR research program of the Universities of Rennes, France.

4. Corpus d'Étude des Langues Vivantes Appliquées à une Spécialité

2 Learner corpora in language learning

The use of corpora in language education has shown its strong potential in three different areas. Many studies have exemplified how learners can benefit from Data Driven Learning (DDL) activities (Boulton, 2017). Corpus exploration activities help learners in their acquisition process by way of observing and manipulating language in authentic contexts. Syllabuses also benefit from corpora. They provide evidence of authentic contexts from which language teaching objectives can be extracted (Granger, 2015). Finally, corpora have also demonstrated a strong potential in iCALL systems (intelligent Computer Assisted Language Learning). Combined with NLP and machine learning methods they can be used to automate language correction(Leacock et al., 2015; Tetreault et al., 2018) in learner productions.

In most studies, most approaches rely on native language corpora. Since they represent the target hypothesis, they are used by learners and automatic systems to evaluate non-native productions. However, in some approaches, including corpus exploration activities, learner dictionary design and error correction, learner corpora are used as sources for learning. But in most cases these approaches focus on errors without highlighting the positive properties of productions (Hawkins and Filipović, 2012).

It is necessary to analyses learner language with more than just error-centric feedback. Some approaches exploit learner corpora to produce meaningful and specific feedback for learners (Shute, 2008) based on positive properties. Some automatic proficiency level prediction methods (Alexopoulou et al., 2013; Pilán, 2018; Ballier et al., 2020) rely on operationalised linguistic complexity metrics (Housen et al., 2012). We endorse the view that complexity metrics can be used as positive properties for the characterisation of learner language. We show that a learner corpus and its data set of metrics can be exploited for the characterisation or the vizualisation of learner writings.

3 Corpus design

3.1 Data collection and task

The corpus includes learner texts in L2 English, German and Spanish collected in two universities of X via a MOODLE Database (see Figure 1) designed specifically for this purpose. The corpus texts were collected during class under the supervision of a language teacher trained on the collection protocol. It includes metadata (Gilquin, 2015; Callies, 2015) about the characteristics of the subjects such as domain of studies, age, number of years studying the L2 and their learning behaviours such as frequency of exposure to L1 and travelling to L1 countries.

In terms of task, the learners were required to conduct two writings. The first one was to describe an experiment/discovery/invention/technology of their choice and the second task was to give their opinion on the impact of the described item. They had 45 minutes in total to complete both tasks.

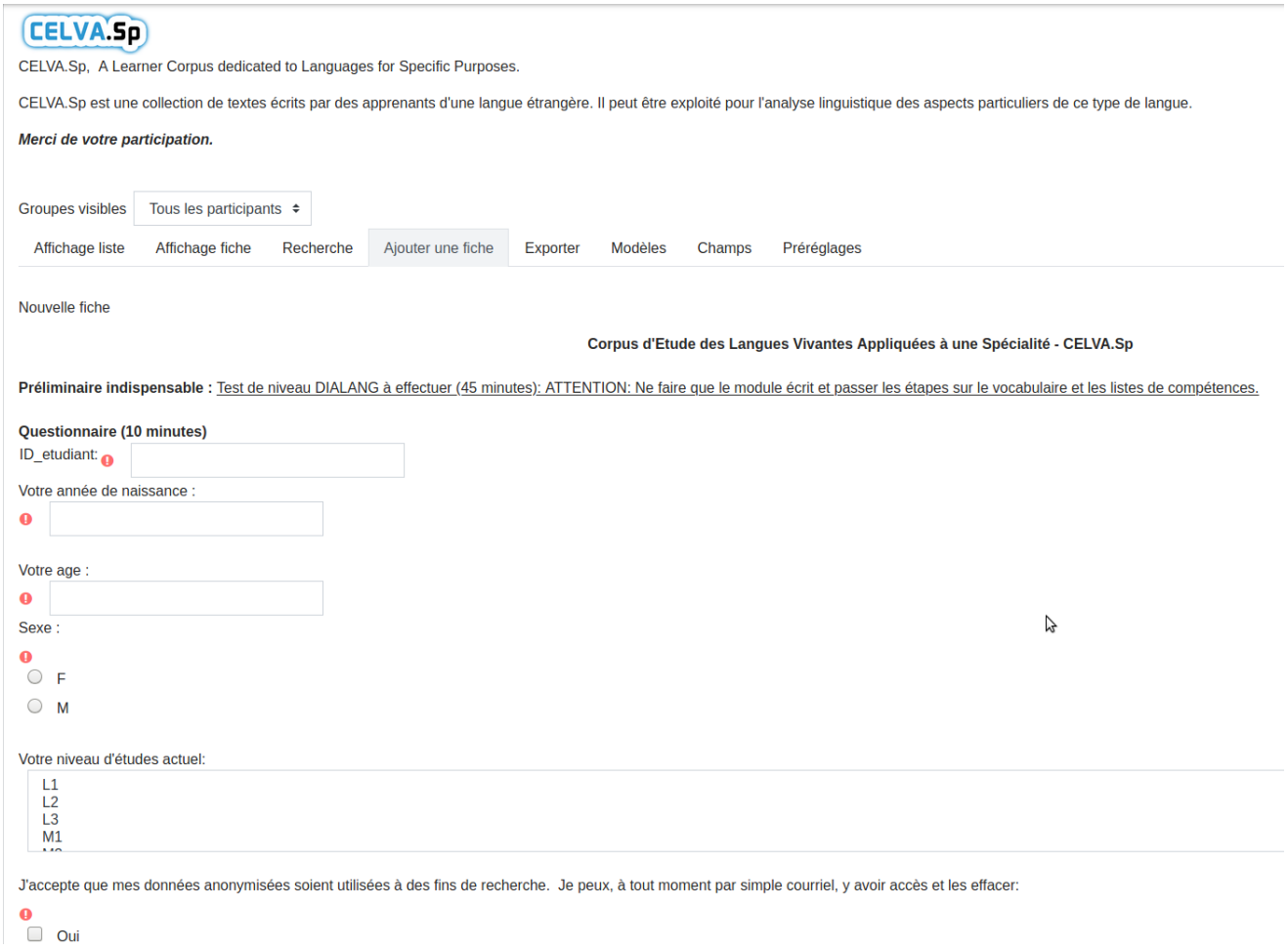

Figure 1. The MOODLE database interface for the collection of the CELVA.Sp corpus

Prior to recording their texts and learner profiles, learners were also requested to carry out the Dialang [5] test (Alderson and Huhta, 2005). For practical reasons, only the written module of the test was used with the exception of the "Placement test" screen and the "Self-assessment- writing" screen. In other terms only the 30 cloze questions were used.

3.2 Annotation

The texts were subsequently annotated in terms of CEFR levels. Two expert teachers and professional language certification examiners annotated the texts and used the latest CEFR written descriptor list (Conseil de l'Europe, 2018, p. 181).

Inter-annotator agreement was operationalised with the Weighted Kappa indicator [6] in order to take the ascending order of the classes into account. It was computed in two phases. Firstly, one random sample of 30 texts was submitted to both examiners (Weighted Cohen's Kappa = 0.65, p-value = 0.000158). Secondly, examiners analysed occurrences of disagreement and a new random sample of 20 texts was added to the first sample. Weighted Cohen's Kappa showed 0.714 agreement (n = 50, p-value = 3.34e-07). In total 235 texts were annotated. Table 1 shows the breakdown in terms of CEFR levels and according to domains of studies and CEFR levels

Table 1. Corpus CELVA.Sp annotated in CEFR levels per domain of study

CEFR levels	A1	A2	B1	B2	C1	C2
Computer science	1	8	22	8	3	1
Medicine	0	8	18	15	3	3
Pharmacy	2	15	29	8	3	6
Biology	17	24	9	1	3	1
Physics	2	8	10	4	1	2
Total	22	63	88	35	13	13

We tested the agreement between the CEFR grades obtained in the Dialang test and the annotated CEFR level of the learners' writings. Inter-annotator agreement showed a 0.616 weighted Cohen's Kappa (n=235 ; p-value = 0) indicating over 60% agreement only.

4 Complexity metrics

The corpus is processed with Natural Language Processing Tools to compute the metrics. L2SCA (Lu, 2014) is used to compute syntactic complexity metrics.

5. see https ://dialangweb.lancaster.ac.uk/
6. we use the irr package in R (R Core Team, 2012)

These metrics rely on syntactic parsing annotation in the form of phrase constituents and parts of speech (POS) produced with Stanford CoreNLP (Manning et al., 2014). By way of pattern matching, specific syntactic items are found and ratios are calculated such as the average number of clauses per sentence or the mean length of sentences. These ratios provide information in terms of sentence complexity, subordination, coordination and particular structures such as compounds and genitives.

Quanteda (Benoit et al., 2018) is an R package used to compute readability and lexical diversity metrics. Readability metrics are based on the morphological features of words to compute different indicator values. The assumption is that indicators operationalise the level of maturity required for reading a specific text and may be indicative of learner proficiency (Lissón, 2017). This includes indicators such as the Coleman Liau, the Dale Chall readability score and the Flesch kincaid grade. They all rely on word length in terms of characters and syllables as well as predetermined lists of words judged as difficult. Lexical diversity metrics provide information on the word frequencies in terms of types and tokens. Several indicators are computed such as Type Token Ratio (TTR), Carroll's Corrected TTR and Yule's K.

A data set of 83 metrics in total has been built and includes 235 observations corresponding to the English subset of the corpus. The data set is also provided with the corpus and its metadata allowing for research in several directions.

5 Companion tools

As well as the MOODLE database and its interface presented in Section 3.1 and made available for download [7], we also created a tool, called VizLing [8], to process all the submitted texts . This allows for easy collection and processing. The tool outputs the metric values for all the submitted texts and creates individual reports in PDF formats that can be handed back to learners.

The reports include different plots presenting the specific metric values obtained by the learners in relation to metric values computed with the CELVA.Sp texts. Figure 2 shows how a learner can analyse a radar chart. The learner's specific metric values are presented in relation to the values of a specific CEFR group. For instance, the Mean Length of T-Unit (MLT) indicator shows that the learner's sentences tend to be longer that those of B1 learners. Conversely the *text.repetitions* indicator shows that the learner tends to repeat words more than B1 learners do. Visualisations show learner profiles on the same plots as the CEFR cohort profiles. Learners can appreciate objective and precise measurements of their productions. It must me stressed that learners need to be guided in reading their reports.

7. see https ://lidile.hypotheses.org/
8. see https ://github.com/LIDILE/VizLing

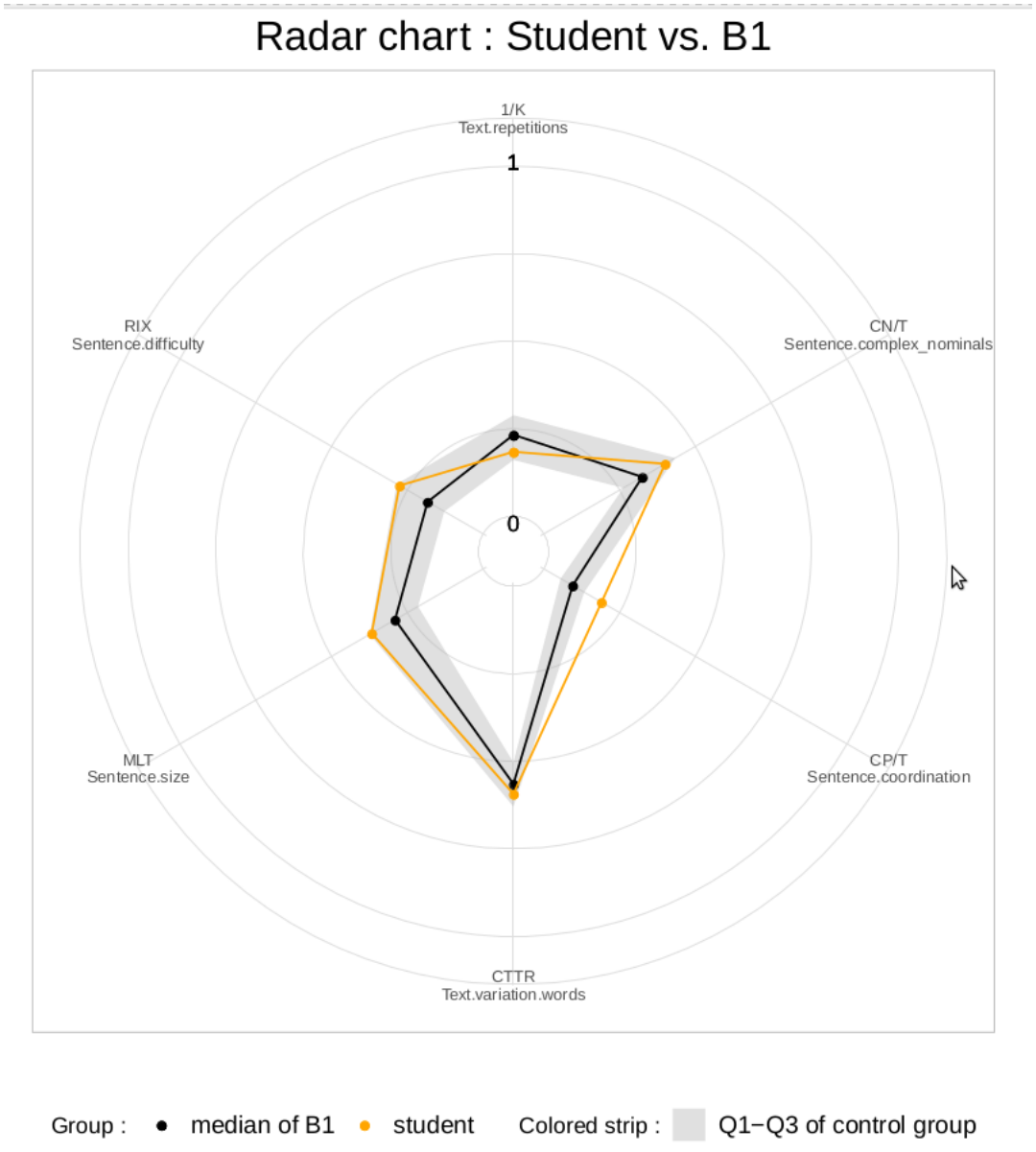

Figure 2. Example of a radar chart in a learner's individual report after processing the text of a learner

6 Conclusion and perspectives

The CELVA.Sp was designed for the study of LSP as it contains writings in Spanish, German and English L2 in several specialised domains. The English subset includes 235 texts written by learners of various scientific domains. This corpus can be exploited with a tool for the automatic analysis of English texts in terms of syntactic and lexical complexity as well as readability. This tool provides visualisations for learners. The corpus is composed of texts, metadata and metric values.

As most complexity metrics are language agnostic, we plan to develop a multilingual version of the tool. After collecting more texts in German and Spanish, it will be possible to compute statistics for different CEFR cohorts supporting comparisons with individuals. By increasing the collected data for each specialised domain, it will also be possible to have reliable statistics cross-referencing CEFR levels, domains and metrics. Learners and teachers will benefit from tools relying on learner corpora to provide them with diagnostics and feedback information.

Bibliographie

Alderson, J. C. and Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing*, 22(3) :301–320.

Alexopoulou, T., Yannakoudakis, H., and Salamoura, A. (2013). Classifying intermediate learner English : a data-driven approach to learner corpora. In Granger, S., Gilquin, G., and Meunier, F., editors, *Twenty years of learner corpus research : looking back, moving ahead*, pages 11–23. Presses Universitaires de Louvain, Belgium.

Ballier, N., Canu, S., Petitjean, C., Gasso, G., Balhana, C., Alexopoulou, T., and Gaillat, T. (2020). Machine learning for learner English. *International Journal of Learner Corpus Research*, 6(1) :72–103.

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Mller, S., and Matsuo, A. (2018). quanteda : An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30) :774.

Boulton, A. (2017). Data-driven learning and language pedagogy. In *Language, Education and Technology : Encyclopedia of Language and Education*, : Encyclopedia of Language and Education.

Callies, M. (2015). Learner Corpus Methodology. In Granger, S., Gilquin, G., and Meunier, F., editors, *The Cambridge Handbook of Learner Corpus Research*, pages 35–56. Cambridge University Press, Cambridge.

Conseil de l'Europe (2018). *Un Cadre Europen Commun de Rfrence pour les Langues : Apprendre, enseigner, valuer Volume complmentaire avec de nouveaux descripteurs*. Conseil de l'Europe, Strasbourg, division des politiques ducatives service de leducation edition.

Gilquin, G. (2015). From design to collection of learner corpora. In Granger, S., Gilquin, G., and Meunier, F., editors, *The Cambridge Handbook of Learner Corpus Research*, number 9-34. Cambridge University Press, Cambridge.

Granger, S. (2015). The contribution of learner corpora to reference and instructional materials design. In Granger, S., Gilquin, G., and Meunier, F., editors, *The Cambridge Handbook of Learner Corpus Research*, pages 485–510. Cambridge University Press, Cambridge.

Hawkins, J. A. and Filipović, L. (2012). *Criterial Features in L2 English : Specifying the Reference Levels of the Common European Framework*, volume 1 of *English Profile Studies*. Cambridge University Press, United Kingdom.

Housen, A., Kuiken, F., and Vedder, I., editors (2012). *Dimensions of L2 performance and proficiency : complexity, accuracy and fluency in SLA*, volume 32 of *Language Learning & Language Teaching (LL&LT)*. John Benjamins Publishing Company, Amsterdam, Pays-Bas, Etats-Unis d'Amrique.

Leacock, C., Chodorow, M., and Tetreault, J. (2015). Automatic grammar- and spell-checking for language learners.

Lissón, P. (2017). Investigating the use of readability metrics to detect differences in written productions of learners : a corpus-based study. *Bellaterra journal of teaching and learning language and literature*, 10(4) :0068–86.

Lu, X. (2014). *Computational Methods for Corpus Annotation and Analysis*. Springer, Dordrecht.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, pages 55–60.

Pilán, I. (2018). *Automatic proficiency level prediction for Intelligent Computer-Assisted Language Learning*. Gothenburg University Publications Electronic Archive, Gothenburg, Sweden.

R Core Team (2012). *R : A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78(1) :153–189.

Tetreault, J., Burstein, J., Kochmar, E., Leacock, C., and Yannakoudakis, H., editors (2018). *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, New Orleans, Louisiana.