LDK 2023

A new learner language data set for the study of English for Specific Purposes at university level

Cyriel Mallart[1], Andrew Simpkin[2], Rémi Venant[3], Nicolas Ballier[4], Bernardo Stearns[5], Jen Yu Li[1], Thomas Gaillat[1]

[1]LIDILE, Université Rennes 2
[2]School of Mathematics, Statistics and Applied Mathematics,University of Galway
[3]LIUM, Université du Mans
[4]CLILLAC-ARP, Université Paris Cité
[5]Insight, Data Science Institute, University of Galway

Abstract

This paper presents the release of a new data set for the study of English as a second language (L2), which is specialised in specific academic domains. The corpus includes 671 texts written by university students of different academic domains. All learners and their CEFR levels had to respond to the same task prompt eliciting language related to a domain. The data set includes structured textual data with rich Universal-Dependency linguistic annotation and metadata. It is available online in the CONLL-U format and can be exploited in several types of NLP tasks related to English L2 analysis.

1 Introduction

This paper reports on the release of the Corpus for the Study of Foreign Languages Applied to a Specialty (CELVA.Sp)[1], a new data set for the study of learner English. Learner corpora have been a topic for research for more than 30 years. They lend themselves to statistical methods for different types of analyses including Contrastive Interlanguage Analysis (CIA), error or linguistic complexity analysis or proficiency assessment. Today, a number of applications rely on learner corpora for modelling tasks. Output models are subsequently exploited in data processing pipelines tuned for specific language learning objectives. Learner corpora have turned out to be an essential resource for Computer-Aided Language Learning (CALL) systems.

In this context, it is essential to use data sets that have been collected with accuracy in controlled environments so as to ensure quality and experimental validity. English learner corpora have benefited from a lot of attention, resulting in the availability of several large corpora such as the Cambridge Learner Corpus (CLC) (Yannakoudakis et al., 2011), the EF™ CAMbridge DATabase (EFCAMDAT) (Geertzen et al., 2013) or the International Corpus of Learner English (ICLE) (Granger et al., 2020). In spite of their sizes, these corpora may suffer from one or more possible limitations such as limited access to raw data files, lack or unclear validity of proficiency annotation, lack of rich behavioural learning metadata. These limitations stem from the fact that learner corpus collection requires a lot of resources in terms of man/woman hours. Collecting such data means identifying learners willing to provide writings or oral recordings together with personal information regarding the learning behaviour, all of this while respecting privacy as required by the European GDPR directive. As a result, access to free, accessible and rich English L2 data sets is not so simple as it may appear. In addition, the aforementioned corpora tend to focus on learners by way of general English writing tasks. As a result, it is difficult to make comparisons between learners of different study domains such as medicine, pharmacy, computer science or sports.

Our proposal is to deliver an English L2 data set designed for the study of L2 English writing skills at university level and across ten different academic domains. We provide writings produced by 671 learners of six levels of proficiency. Learners' metadata are included and inform researchers on the learners' backgrounds and their behaviour

[1]Corpus d'Etude des Langues Vivantes Appliquées à une Spécialité. Available from the Huma-Num Nakala repository located at https://nakala.fr/10.34847/nkl.41d57kb0, DOI 10.34847/nkl.41d57kb0

in learning English, e.g. exposure to English media, reading attitude, language trips and secondary school focus on advanced English classes. This data set is available in an interoperable format allowing automatic processing methods.

2 Related work

A number of learner English writing corpora exist on the commercial market. The International Corpus of Learner English (ICLE) version 3 is certainly one of the main resources in this field. It includes 9,529 long essays written by learners of twenty-six L1s and associated with educational metadata. It is also possible to apply for a non-commercial user licence for access to its exploration interface. The Cambridge Learner Corpus is commercial in its full version, but it includes a publicly released subset made up of exam scripts taken by candidates of the First Certificate in English (FCE). This subset includes 1,244 scripts together with proficiency marks and error annotation but it lacks metadata concerning the exam takers.

In the realm of non-commercial data corpora, the EFCAMDAT corpus is a collection of learner writings which have been classified in terms of proficiency levels. Its 1,180,309 scripts make it the biggest learner corpus of its kind as far as we know. It comes with some learner metadata such as learner nationality, EF™ proficiency levels, lesson units, task topics and grades. The learners' backgrounds are unknown and the evaluation of proficiency annotation is not reported in the paper.

Some learner corpora specifically focus on university students. The University of Pittsburgh English Language Institute Corpus (PELIC) (Juffs et al., 2020) focuses on university students and provides 46,230 scripts split into many different generic writing task topics. The NUS Corpus of Learner English (NUCLE) (Dahlmeier et al., 2013) is made up of about 1,400 essays, including error annotation, written by university students. Likewise the ASAG corpus (Tack et al., 2017) provides short texts written by third-level students as short answers to general-English questions. The corpus includes a subset of 299 writings that were graded according to the CEFR levels.

The aforementioned corpora rely on data that come from learners of English of unknown academic fields. The writing prompts were designed to fit all possible types of students and thus were not necessarily linked to the field of studies. Yet, at university level, there is a need to study how learners of English for Specific Purposes (ESP) construct their linguistic knowledge in relation to their future professional domain. In this respect, the Varieties of English for Specific Purposes dAtabase (VESPA) (Paquot et al., 2022) provides more than 900 long essays written by learners of different L1 and different academic fields. This type of data is very useful to help explore and compare learner linguistic profiles across several domains.

We propose a more modest ESP corpus. Its main difference is that it relies on a single prompt designed to elicit domain-specific writings of the same genre and discourse types. This allows for comparisons between the writings of students of different academic fields. The texts are 200- to 300-words long and reflect a typical writing requirement set by language teachers in class. In addition, the writings are associated with learning-behaviour metadata and learner proficiency.

3 Corpus design

3.1 Data collection and task

The corpus includes learner texts in L2 English collected in two French universities of the same city. The learners were mostly French students between 2018 and 2020 at undergraduate level, ranging from first to third year.

The data was collected via a MOODLE Database[2] (Dougiamas and Taylor, 2003) designed specifically for this purpose. It can be installed on any MOODLE server for further collection in other educational environments.

The corpus texts were collected during class under the supervision of university language teachers trained on the collection protocol. It includes recommended 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 L2 and travelling to L2 countries. Database fields were defined to control the possible values that could be entered, hence avoiding too much variability in categorical data names. The corpus data were then be exported as a UTF8 .csv file for further processing.

In terms of task, the learners were required to conduct a writing task with one and the same

[2]The MOODLE package is available from Gitlab URL

prompt. It required the description of an experiment/discovery/invention/technology/technique of their domain followed by their opinion on the impact of the described concept. The prompt was chosen as it allowed each learner to elaborate text dedicated to their own domain while ensuring the same text genre and discourse type. The learners had 45 minutes to complete the task.

Prior to recording their texts and learner profiles, learners were also requested to carry out the Dialang[3] test (Alderson and Huhta, 2005). For practical reasons related to test taking duration in class, 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 Data cleaning

After collecting the data, some records were discarded. These include the records where no email address is known, which is due to database tests. Duplicates, that is, records that contain exactly the same text from the same student but at two different times, were reduced to a single occurrence with the earliest date set as the submission date. Finally, we removed records in which the student wrote in Spanish or German while declaring that their L2 was English, and the samples in which the text was shorter than 10 words.

Some records were cleaned. The texts written by the students were cleared of all HTML formatting, while conserving the original paragraph structure. We simplified a variable that previously contained the names of advanced language sections followed by a student into a binary one. It now stores whether the student followed an advanced language curriculum in the past or not. Dates were set to a uniform format.

3.3 Data pseudonymization

In order to comply with the GDPR guidelines, the data were pseudonimized and learner-identifying information removed. Identifying information covers name, email address, age and level of studies. Other metadata relevant to the learning behaviour, and that do not allow for identification of an individual student, were kept, such as L1, number of years studying the L2, reading frequency, exposure to the language or number of trips taken in an English-speaking country. Learners who answered negatively to whether they consented to the use or distribution of their data were also removed.

Each learner is represented by a secure encoding of their email address, created through an HMAC algorithm (Bellare et al., 1996) that uses a SHA256 cryptographic hash function. This algorithm encodes the email address of the student to a unique 64 letters and digits long pseudonym. This choice ensures unicity of the pseudonym. A secure SHA256 encoding of the email address requires a secret key, known only to the curators of the data set. Indeed, one pseudonym represents one student only. This will allow following the progression of a given student across time or tasks in the future. Should a participant revoke their consent to having their data used, the curators of the data set are capable of finding the records of this individual to remove them from the data set. This complies with the GDPR's guidelines on the right to request the rectification (and erasure) of personal data.

Beyond the metadata, learners may also disclose personal information in their writings. We replaced names with a placeholder, "Alex Dupont", instead of other methods such as initials or special symbols in order to stay as faithful as possible to the original language used by the student.

3.4 Linguistic annotation

In addition to plain text, the data set also contains linguistic information relying on the framework of Universal Dependencies (de Marneffe et al., 2021). The annotations notably include Universal Dependency tagged part-of-speech, lemmas of tokens, and morphological features such as case, number, gender, etc. These were obtained with the UDPipe pipeline (Straka et al., 2016) using the English model trained on the GUM corpus[4] (Zeldes, 2017) as it was shown to be very reliable for POS and dependency annotation on L1 and L2 (Kyle et al., 2022). Evaluation of annotation accuracy was not conducted on these data .

4 Data set description

4.1 Metadata and text descriptions

The data set includes 671 writings from French-L1 learners and made up of 215 words on average (SD = 116.35) as shown in Figure 1. The writings are spread over ten different academic fields taught in

[3] see https://dialangweb.lancaster.ac.uk/

[4] english-gum-ud-2.5-191206

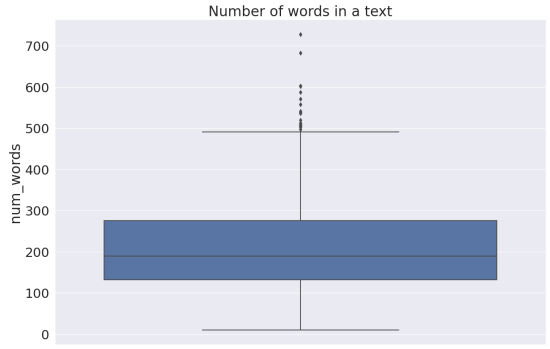

Figure 1: Distribution of the number of words per text

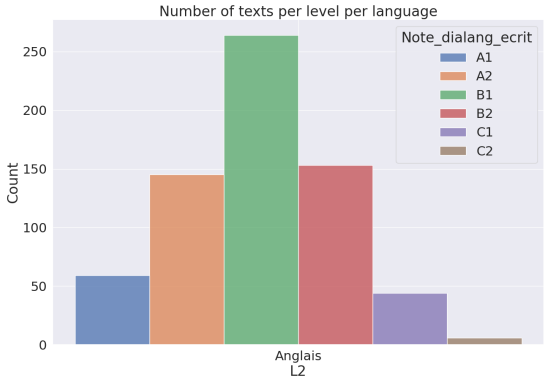

Figure 2: Distribution of the number of texts per CEFR level

the universities of the city. Table 1 provides a detailed view of the data. Note that the imbalance is due to the domains in which data-collecting teachers were involved.

Domains	**texts**
Media Studies	199
Earth and Life Sciences	109
Medicine	96
Pharmacy	82
Computer Science and Electronics	65
Physics and Chemistry	40
Education Sciences	38
Science and Technology of Sport and Exercise	38
Mathematics	2
Social Sciences and Humanities	2

Table 1: Distribution of the number of texts per academic domain

All the writings are linked to the CEFR levels obtained by the learners in the DIALANG test. Figure 2 shows the distribution of texts per CEFR level. Interestingly, the number of words increases as CEFR levels increase except for the top C2 level. C2 learners seem to deflate their writing volume, maybe in favour of better pragmatic efficacy in discourse complexity and coherence. Figure 3 shows the variations of the number of words per level, giving an insight into the writing productivity of the learners. The metadata and the texts are all included in the same CSV file. The linguistic information about all the textual elements is included in a separate data file as described in Section 4.2. Both files are indexed with the pseudonymized identifier as described in Section 3.3.

4.2 Data formats

The data set adopts the CONLL-U format as part of a CSV file. More specifically, each CONLL-U representation is formatted as a string, and for each text a single string is stored in the *conllu_text* column of the CSV. In this format each text is associated with a multi-layer representation of its linguistic annotation. For instance, each token is assigned the following information:

- FORM,
- LEMMA,
- UPOS,
- XPOS,
- FEATS (List of morphological features),
- HEAD (Head of the word dependency governor),
- DEPREL (Universal dependency relation to the HEAD),
- DEPS (A list of head-dependency relations pairs),
- MISC (Any other annotation such as givenness)[5].

Thanks to the encoded dependency information, the files can subsequently be visualized with the CoNLL-U Viewer[6] or queried with tools such as Grew-match (Amblard et al., 2022).

In addition, we added the metadata to the files. The metadata are accounted for with categorical and numerical variables named in French. They are:

- *Nb_annees_L2*: Number of years studying L2 English
- *L1*: Native language
- *Domaine_de_specialite*: Academic domain of the learner
- *Sejours_duree_semaines*: Total number of weeks spent in English speaking countries

[5] See https://universaldependencies.org/format.html for detailed information

[6] Available at https://universaldependencies.org/conllu_viewer.html

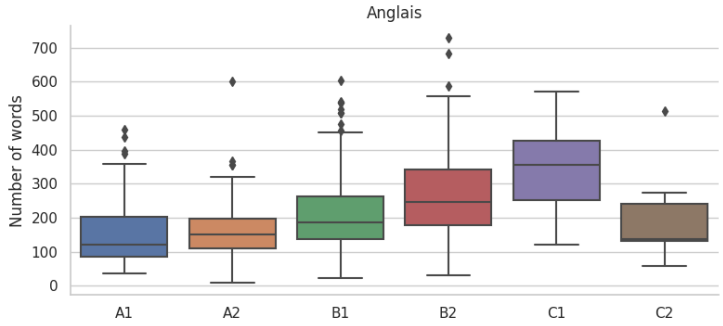

Figure 3: Distributions of texts according to their number of words and the CEFR levels of the learners

- *Sejours_frequence*: Number of trips
- *Lang_exposition*: Out-of-class exposure to L2 English (movies, radio ...)
- *Note_dialang_ecrit*: CEFR class with the DIALANG test
- *Lecture_regularite*: Reading frequency (daily, weekly, montly)
- *autre_langue*: Other L2 being learnt
- *tache_ecrit*: Identifier of writing task (only one)
- *Texte_etudiant*: Texts written by students
- *Date_ajout*: Date of writing
- *pseudo*: Pseudonymised ID of learner

5 Exploitation of the data set

This data set may be exploited in a wide array of tasks. ESP corpora play an important role in the field of academic language research as they help identify L2 developmental patterns linked to a specialised domain. They can thus support course material design with adapted content depending on academic profiles. Such data are useful for the design of Intelligent Computer-assisted Language Learning (ICALL) systems. These systems rely on supervised learning approaches that use learner corpora for error detection (Tetreault et al., 2018) or CEFR classification (Yannakoudakis et al., 2018; Gaillat et al., 2021) or language feature visualization (Gaillat et al., 2023).

Researchers involved in the ESP field will find the corpus useful for linguistic exploration and its potential for multidimensional analysis combining learning behaviour information with fine-grained linguistic annotation. In this respect, the CELVA.Sp data set can be exploited with a the Grew-match tool which provides for linguistic queries. Note that, thanks to the data and metadata formats, it is possible to sub-sample the data in order to obtain balanced datasets.

The data set could also be used in supervised learning tasks as it offers well-structured data. Traditional methods of machine learning such as logistic regression, support vector machines, random forests or gradient tree boosting require a large amount of tabular data. The CELVA.Sp data set provides tabular metadata, with little work required to create either tabular bag-of-word (Harris, 1954) features from the raw text or more complex dependency or morphological features from the linguistic annotations. More recent deep learning methods, such as convolutional neural networks (Kim, 2014), recurrent neural networks (LeCun et al., 2015) and transformer-based neural networks (including BERT (Devlin et al., 2019) and chatGPT[7]), require an unprecedented amount of data to train. However, the power of these models lies in the fact that they can be pre-trained on vast amounts of unannotated data from various sources, and then fine-tuned on a precise natural language task using task-relevant data. (Zhang et al., 2021) trained a BERT model on a task of textual entailment using the RTE dataset (Dagan et al., 2006) which consists of only 2,500 training data samples. The model achieved a 69.5 F1 score without any optimization. Our data set fits within this paradigm, with enough annotated learner data to fine-tune state-of-the-art deep learning models and leverage the predicting power of those models for tasks such as CEFR level prediction, or error modelling.

We intend to exploit this corpus as part of a Computer-Assisted Language Learning (CALL) system dedicated to the automatic analysis of learner language at university level. The corpus will be used to model learner proficiency across different academic domains. The system will display linguistic feature visualizations within the

[7]https://openai.com/blog/chatgpt

MOODLE system.

Further data enrichment is also planned. The corpus texts will be annotated by six language-certification experts following CEFR guidelines and inter-rater agreement will be evaluated. The final corpus will include texts of other L2s than English, including German, Swedish and Spanish. Keylog information recorded at time of writing will also be included. More writing tasks will be added for learners of all levels to ensure genre variety. The corpus will be available online.

6 Credits

We wish to thank all the language teachers who helped in collecting the data. This projected is funded by the French National Research Agency ANR-22-CE38-0015-01

References

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

Maxime Amblard, Bruno Guillaume, Siyana Pavlova, and Guy Perrier. 2022. Graph querying for semantic annotations. In *Proceedings of the 18th Joint ACL - ISO Workshop on Interoperable Semantic Annotation within LREC2022*, pages 95–101. European Language Resources Association.

Mihir Bellare, Ran Canetti, and Hugo Krawczyk. 1996. Keying Hash Functions for Message Authentication. In *Advances in Cryptology*, pages 1–15. Springer.

Marcus Callies. 2015. Learner corpus methodology. In *The Cambridge Handbook of Learner Corpus Research*, Cambridge Handbooks in Language and Linguistics, pages 35–56. Cambridge University Press.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. volume 3944, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg. Book Title: Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment Series Title: Lecture Notes in Computer Science.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a Large Annotated Corpus of Learner English: The NUS Corpus of Learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics.

Martin Dougiamas and Peter Taylor. 2003. Moodle: Using Learning Communities to Create an Open Source Course Management System. In *Proceedings of the EDMEDIA 2003 Conference*, pages 171–178. Association for the Advancement of Computing in Education.

Thomas Gaillat, Antoine Lafontaine, and Anas Knefati. 2023. Visualizing Linguistic Complexity and Proficiency in Learner English Writings. *CALICO Journal*, 40(2):178–197.

Thomas Gaillat, Andrew Simpkin, Nicolas Ballier, Bernardo Stearns, Annanda Sousa, Manon Bouyé, and Manel Zarrouk. 2021. Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach. *ReCALL*, 34(2). Publisher: Cambridge University Press.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2013. Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat). In *Selected Proceedings of the 2012 Second Language Research Forum: Building Bridges between Disciplines*. Cascadilla Press.

Gaëtanelle Gilquin. 2015. From design to collection of learner corpora. In *The Cambridge Handbook of Learner Corpus Research*, Cambridge Handbooks in Language and Linguistics, pages 9–34. Cambridge University Press.

Sylviane Granger, Maïté Dupont, Fanny Meunier, Hubert Naets, and Magali Paquot. 2020. *The International Corpus of Learner English. Version 3*. Presses universitaires de Louvain.

Zellig S. Harris. 1954. Distributional structure. *WORD*, 10(2-3):146–162.

Alan Juffs, Na-Rae Han, and Ben Naismith. 2020. The University of Pittsburgh English Language Institute Corpus (PELIC).

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics.

Kristopher Kyle, Masaki Eguchi, Aaron Miller, and Theodore Sither. 2022. A Dependency Treebank of Spoken Second Language English. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 39–45, Seattle, Washington. Association for Computational Linguistics.

Yann LeCun, Y. Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521:436–44.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, (2):255–308.

Magali Paquot, Tove Larsson, Hilde Hasselgård, Signe O. Ebeling, Damien De Meyere, Larry Valentin, Natalia J. Laso, Isabel Verdaguer, and Sanne van Vuuren. 2022. The Varieties of English for Specific Purposes dAtabase (VESPA): Towards a multi-L1 and multi-register learner corpus of disciplinary writing. *Research in Corpus Linguistics*, 10(2):1–15.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 4290–4297. European Language Resources Association.

Anaïs Tack, Thomas François, Sophie Roekhaut, and Cédrick Fairon. 2017. Human and Automated CEFR-based Grading of Short Answers. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 169–179. Association for Computational Linguistics.

Joel Tetreault, Jill Burstein, Ekaterina Kochmar, Claudia Leacock, and Helen Yannakoudakis, editors. 2018. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics, New Orleans, Louisiana.

H. Yannakoudakis, Øe Andersen, A. Geranpayeh, T. Briscoe, and D. Nicholls. 2018. Developing an automated writing placement system for ESL learners. Accepted: 2019-02-16T00:31:04Z Publisher: Informa UK Limited.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 180–189. Association for Computational Linguistics.

Amir Zeldes. 2017. The gum corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. 2021. Revisiting Few-sample BERT Fine-tuning.