Leonardo Zilio, Rodrigo Wilkens & Cédrick Fairon

Université Catholique de Louvain, Louvain-la-Neuve, Belgium

{leonardo.zilio,rodrigo.wilkens,cedrick.fairon}@uclouvain.be

Analyzing grammatical structures in texts written by language learners

Bio data

Leonardo Zilio is a postdoctoral researcher at the Center for Natural Language Processing (CENTAL), Université catholique de Louvain (UCL), Belgium, where he develops project SMILLE. He obtained his PhD in Linguistics at the Federal University of Rio Grande do Sul (UFRGS), Brazil, and carried out a one-year doctoral stay at the Laboratoire d'Informa-tique de Grenoble (LIG), France. He is also translator of German, English and Portuguese.

Rodrigo Wilkens is a postdoctoral researcher at the Center for Natural Language Processing (CENTAL), Université catholique de Louvain (UCL), Belgium. He has a master's and PhD degree in Computer Science by the Federal University of Rio Grande do Sul (UFRGS), Brazil, and carried out a doctoral stay in the Massachusetts Institute of Technology (MIT), USA.

Cédrick Fairon is professor at the Université catholique de Louvain (UCL), Belgium, and director of the Center for Natural Language Processing (CENTAL). After a PhD in Computer Sciences at the University Paris 7, he was postdoctoral researcher at New York University (NYU, Linguistic Department) and worked as Associate Research Scientist at Educational Testing Service (Princeton, NJ) on automatic test generation.

Abstract

In Secon Language Acquisition (SLA), the evaluation of a language learner's proficiency is a task that normally involves comparing the learner's production with a learning framework of the target language. One of the most well known frameworks is the Common European Framework for Languages, which establishes communicative goals for language learning in general. In this study, we automatically annotated a corpus of texts produced by language learners with pedagogically relevant grammatical structures and observed how these structures are being employed by learners from different proficiency levels. We analyzed the use of structures both in terms of evolution along the levels and in terms of level in which the structures are used the most. The annotated resource presents a rich source of information for teachers that wish to compare the production of their students with those of already certified language learners.

Conference paper

Introduction

The evaluation of a language learner's performance when producing texts in a foreign language is not an easy task. The factors that impact the overall categorization of texts produced by learners are many, roughly ranging from vocabulary, to syntax and semantics, and to discourse strategies. One way of facilitating this task is to have a profile of the target skills, so that there are some hints on what to expect from a learner in each language level.

For this reason, there are different frameworks that organize how the different language skills should be targeted at each step, while also indicating the required skills for the evaluation of a learner's proficiency. Examples of this type of frameworks are the Common European Framework for Languages (CEFR) and the Cambridge ESOL, which are based on levels, and IELTS and TOEFL, which are based on scores. These frameworks pinpoint, in differently organized fashion, how it is expected that the second language learning will take place for the learner, by listing skills and associating them with an expected level (or score).

In this study, we use a different approach. Instead of pointing out the skills that the learner should be able to master in order to achieve a certain level of proficiency, our objective is to look directly at the production of learners that have already achieved a certain degree of proficiency. By investigating texts produced by learners and quantitatively observing how different types of grammatical structures are used by learners from different language levels and by analyzing the distribution of grammatical structures in their textual production, we aim at finding out which structures are more or less active and how they evolve in terms of use along the different language mastery levels. This generates a language profile that presents some anchor points that teachers can use for evaluating learners, for instance, by comparing the use of specific grammatical structures in text produced in classroom with the average of texts produced by language learners in the same level.

For describing the distribution of grammatical structures in texts produced by language learners, we annotated an SLA corpus with pedagogically relevant grammatical structures, which are referred to in, for instance, learner's grammars and the English Grammar Profile (EGP). Our resource, which was named *SLA in Grammatically Annotated Texts* (SGATe)[30], contains more fine-grained information than it would be possible to automatically retrieve from common parsing methods. As such, SGATe provides teachers with the possibility of comparing the written production of their own students with a corpus that displays the use of grammatical structures by certified learners.

Related Work

Among the different frameworks that describe how a second language should be learned, for this study, the CEFR is of special relevance. The CEFR (Verhelst et al., 2009) presents a guide in terms of language levels and content that is meant to serve as a parameter for second language learning in the European Union. It provides a description of communicative goals that a learner should achieve in each of six main levels: A1, A2, B1, B2, C1, and C2. As a general guide that was not designed to cover specific languages, but to present broad communicative guidelines, it leaves various gray areas in terms of the learning process, so that the information for each level does not cover the different needs of a language learner regarding specific grammar and vocabulary content, or even a specific language. As such, the curricula of different language courses do not need to be necessarily the same even if they follow the specified CEFR levels (Alderson, 2007; Little, 2007).

Since we intend to observe the distribution of grammatical structures in a corpus of written production, it is also important to consider systems that were developed for annotating pedagogically relevant information. In this regard, we employed the SMILLE system, which annotate grammatical structures based on the CEFR. The SMILLE system (Zilio and Fairon, 2017; Zilio, Wilkens and Fairon, 2017a; 2017b) recognizes 107 pedagogically relevant grammatical structures in texts. These structures were derived from the pedagogical curriculum for English of Altissia's online learning platform based on the CEFR, so that they are organized by language level. The SMILLE system uses the Stanford Parser (Manning et al., 2014) as basis and then retrieves more complex, pedagogically relevant grammatical information by means of a set of rules that go beyond parser information. These include

30 SGATe is available at http://cental.uclouvain.be/resources/smalla_smille/sgate/.

structures such as *verb + infinitive with "to"*, *connectives of time*, *quantifiers*, *conditionals*, *gerunds as verb complement*, *"let's" + infinitive*, etc.

Methodology

For being able to describe how language learners actually use the grammatical structures they learn, we selected the EF-Cambridge Open Language Database (EFCAMDAT) (Geertzen, Alexopoulou and Korhonen, 2013), which presents a collection of texts produced by learners of English from different levels of proficiency. The corpus amounts to more than 83 million words and is divided according to the Common European Framework of Reference for languages (CEFR). Each document in the corpus has a score indicating how well the learner performed in the task and is linked to a specific topic (e.g. *introducing yourself by email*).

We used the SMILLE system to annotate the corpus with 107 grammatical structures that are pedagogically relevant and analyzed their distribution on the different language levels. These grammatically rich annotations that were added to the EFCAMDAT corpus gave birth to SGATe, namely *SLA in Grammatically Annotated Texts,* a resource in which it is possible to observe the second language acquisition of different learners in terms of pedagogically relevant grammatical structures. Since many of these structures are complex and require rules on top of parser information for being annotated, and since the automatic annotation was not conducted on texts produced by native speakers of English, we conducted an evaluation in terms of precision.

This evaluation was carried out by one linguist and was designed to verify how well SMILLE's handcrafted rules can annotate a corpus of texts produced by learners, including even the most basic levels. Since the structures are automatically annotated, the evaluation also contributes to show on which of the annotated structures we can rely for analyzing the annotated data. The evaluation was carried out on a random sample of the corpus: we extracted a sample of 40 documents from each of the 6 CEFR levels, totaling 240 documents.

Since it would be impossible to evaluate every grammatical structure that was annotated by SMILLE's system in SGATe, and since some of them simply rely on the parser's morphosyntactic annotation, we selected 33 structures to be evaluated that do not rely on morphosyntactic annotation[31] and that present a general overview of the features that SMILLE can annotate[32].

Evaluation

We excluded from the results those annotation errors that were caused by bad spelling or structural organization of sentences, but these were a minor issue, representing only 1.24% of the sample. The overall precision of the system for the evaluated structures was 90.10% (weighed precision: 92.46%), with median at 97.50%. As shown in Table 1, when we look at the differences from level to level, we see a bad overall precision at level A1, and then no palpable difference between the other levels, but a much higher overall precision score. There are many possible reasons for this discrepancy from the A1 level to the others, one of them is the possibility that A1 documents present a writing style that lacks naturalness, and this makes it harder for the parser and for the system's rules to recognize the correct text patterns required for annotating the grammatical structures.

Table 2 shows the precision scores for the different evaluated structures in our sample of SGATe. Although most of the structures had a good precision overall, some structures had performance below 60%, like gerunds as subject of verb, that had an overall precision of

31 There is only one structure that is based on morphosyntactic annotation, and that is the genitive marker. We included this annotation, because it is an important grammatical structure of the English language, and sometimes it poses a problem for learners.

32 A list of the features, with examples, is described by Zilio, Wilkens and Fairon (2018).

55.56%, and did not perform well in any level. Other structures, like imperatives, had lower performance overall (82.03%), but performed very well if we exclude the A1 and A2 Levels (90.48%). The same is true for the genitive marker, which had a low performance in Levels A1 through B1, which pulled its overall performance down to 67.53%, but actually got a high score in the more advanced levels (90.20%). We verified a similar result regarding connectives, but this time in terms of granularity. The evaluation of connectives showed an overall precision of 87.06%, but we also observed that two classes of connectives, namely connectives of example and connectives of purpose, had a much lower precision score (58.02% and 63.36%, respectively), which was compensated by the other classes. This precision evaluation served to show us where the weaknesses and streghts of the annotation lie, so that we could proceed with a deeper analysis of SGATe for describing a language learning profile.

Table 1 - Precision scores for each level

CEFR levels	Overall precision (%)	Weighed precision (%)
A1	58.54	67.21
A2	89.84	91.05
B1	91.75	91.40
B2	90.89	91.70
C1	91.02	90.48
C2	90.81	90.65

Profile of Grammatical Structures

The SGATe (SLA in Grammatically Annotated Texts) resource comprises the entire EF-Cambridge Open Language Database (EFCAMDAT) annotated with pedagogically relevant grammatical structures. For a first exploratory analysis, we looked at the distribution of structures along the different CEFR levels, analyzing the increase or decrease in their occurrence. For that, we used a linear regression algorithm that shows the tendency of structure use across the levels. In terms of selection of grammatical structures for this observation, we analyzed verb tenses and other structures that depend only on the parser's morphosyntactic information, and, from the structures that we evaluated in this study, we selected only those that had precision scores above 90%.

As we observed in the Evaluation section, the automatic annotation doesn't perform well in Level A1, so, for the profiling presented here, we excluded data from that level. We also filtered out documents from the corpus for which the score was lower than 90%[33], because texts with lower scores may present some errors that can interfere with the automatic annotation. We also balanced as best as we could the number of texts from each level, so that Levels A2, B1 and B2 had 9 thousand documents each, and C1 had more than 4 thousand documents[34]. As a final step, we normalized the frequency of the grammatical structures in each level by using a frequency-per-sentence score, which was further converted to logarithm, to compensate for the fact that language data tend to appear in a Zipf distribution.

After this balancing and normalization process that was performed on SGATe data to give us a more reliable information on the tendency of use of grammatical structures along the levels, we divided the structures in three categories, regarding their tendency to evolve along the levels: increasing tendency (angle of the line above 30 degrees), decreasing

33 This is based on the actual score that was given to the texts by L2 evaluators of the Cambridge University while assessing the learner's performance on an exam.

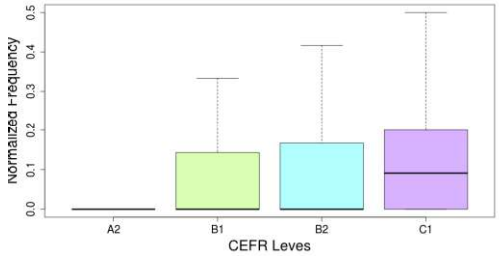
34 All documents with scores above 90% were included in the C1 data. C2 documents were too few to include.

tendency (angle of the line below -30 degrees) and neutral tendency (angle of the line between -30 and 30 degrees). As a means of ensuring the reliability of our results, we considered the linear tendency reliable only if the error of the slope in the linear model scored below 0.15.

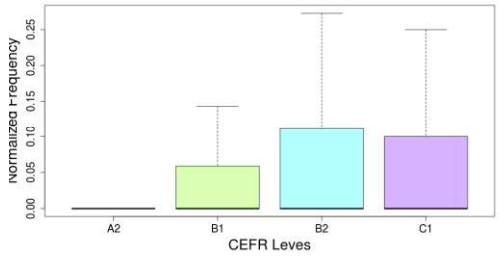
Table 2 - Precision scores of evaluated structures

#	Structure	Total	Precision
1	Gerunds after preposition	109	95.41%
2	Gerunds as complement	3	100.00%
3	Gerunds (no change of meaning)	18	100.00%
4	Gerunds (change of meaning)	2	100.00%
5	Gerunds as subject of a verb	9	55.56%
6	Adjective + infinitive with "to"	93	94.62%
7	Noun + infinitive with "to"	75	74.67%
8	Verb + infinitive with "to"	370	91,35%
9	"let"/"make" + infinitive	25	88.00%
10	"let's" + infinitive	2	100.00%
11	"rather"/"better" + infinitive	1	100,00%
12	Present perfect continuous	12	100.00%
13	Past perfect continuous	2	100.00%
14	Future perfect	3	100.00%
15	Imperatives	128	82.03%
16	Passive voice	233	87.12%
17	Passive adverbs	29	72.41%
18	Connectives	765	87,06%
19	Relative clauses	103	94,17%
20	First conditional	38	100.00%
21	Second conditional	12	100.00%
22	Third conditional	1	100.00%
23	Hypothesis: "would"	48	97.92%
24	Hypothesis: "would have"	1	100.00%
25	Prepositional verbs	199	97.49%
26	Phrasal verbs	104	96.15%
27	"wish" followed by past	1	100,00%
28	Genitive marker	77	67.53%
29	Quantifiers	320	97,50%
30	Special forms of plural	109	99.08%
31	Semi-auxiliaries	79	75,95%
32	Question tags	3	100,00%
33	Wh-questions	35	97,14%

We present here the structures divided by category with information about the angle in brackets. These are the structures with an **increasing tendency**: *adjectives followed by infinitive with "to"* (53°), *relative clauses* (53°), *gerunds after preposition* (52°), *past perfect tense* (51°), *passive voice* (45°), and *gerunds as complement of a verb* (34°). Two examples of these structures are plotted in Figure 1. These are the structures that tend to be roughly **equally** used along the levels A2 to C1: *imperatives* (-3°), *verbs "let" or "make" + infinitive without "to"* (13°), *present participles* (3°), and *present simple* (-1°). Two examples of these structures can be seen in Figure 2. Finally, these are the structures that presented a **decreasing tendency**: *short forms* (-43°), *present continuous* (-43°), and *gerunds instead of infinitive (no change of meaning)* (-35°). We plotted two examples of these structures in Figure 3.

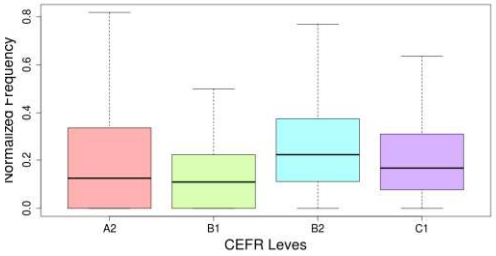


1a. Passive Voice

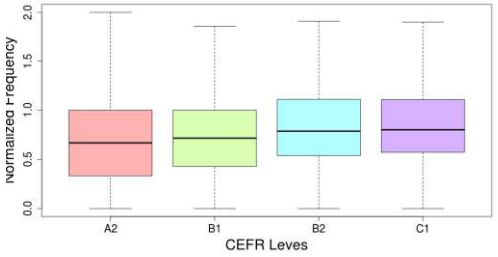


1b. Relative Clauses

Fig. 1 – Examples of structures that have a tendency
to be progressively more prominent along the language levels.

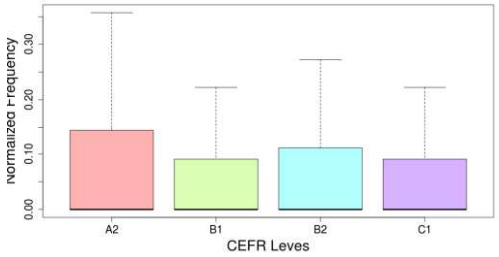


2a. Present Participle

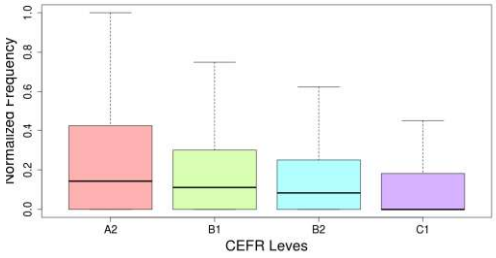


2b. Present Simple

Fig. 2 – Examples of structures that have a tendency
to be equally used along the language levels.



3a. Present Continuous



3b. Short Forms

Fig. 3 – Examples of structures that have a tendency
to be progressively less prominent along the language levels.

The tendency lines presented some interesting information, like the decrease in the use of short forms and the present continuous, while the past perfect and relative clauses get more used. Passive voice also has an increasing tendency, which is expected, because it is

considered to be a more complicated structure to master. This type of information may aid a language teacher to evaluate the progression of learners, by matching their production in terms of grammatical structures to a more ample sample of texts.

Since many structures do not present a clear ascending, descending or neutral tendency (i.e., the error of the slope was 0.15 or higher), we also looked at the peaks of use of each grammatical structure. We used the same data that was normalized by sentence, and we observed in which levels the structures occurred the most (considering a confidence interval of 95% for determining if the difference was significant). Structures that occurred the most at Level A2 were the following: *short forms, past simple, past simple of the verb "to be", past simple of the verb "to have", past continuous, present simple of the verb "to do", present continuous*, and *gerunds instead of infinitive (no change of meaning)*. These are the structures that occurred the most at Level B1: *use of "going to", future perfect, future*, and *expression "let's" followed by infinitive*. Structures that occurred the most at Level B2 were the following: *genitive markers, present participles*, and *present simple of the verb "to be"*. These are the structures that occurred the most at Level C1: *first conditional, second conditional, hypothesis with "would", future continuous, gerunds after preposition, imperatives, passive voice, past perfect, past perfect continuous, present perfect of the verbs "to be" and "to have", present perfect continuous, present perfect, relative clauses, verbs "let" or "make" + infinitive without "to", adjective + infinitive with "to", verb + infinitive with "to"*, and *connectives*.

With this second analysis, we could observe that the verb tenses are well distributed along the corpus: present simple and present continuous, and past simple of auxiliary verbs at level A2, followed by future at Level B1, and then the perfect tenses at Level C1. Connectives are also more concentrated at Level C1, which was a bit of a surprise, since, for instance, the English Grammar Profile tends to present them as lower level structures. The same is true for first and second conditionals, which are normally regarded as A2 or B1-level structures, but have a greater concentration at level C1. This is maybe a sign of the difference between the time of learning and the actual mastery of the grammatical structure.

Conclusion

In this paper, we described the automatic annotation of the EF-Cambridge Open Language Database (EFCAMDAT), a corpus of texts produced by language learners, with pedagogically relevant grammatical information. This layer of annotation that was added to the EFCAMDAT originated SGATe (SLA in Grammatically Annotated Texts) and that allowed us to analyze the distribution of grammatical structures in the production of language learners. As such, we could describe the actual active use of structures by the learners. This same annotation of SMILLE could be employed by language teachers to their learners' texts in order to support an analysis of how they are grammatically structured, facilitating text evaluation. It is important to highlight that the structures annotated by SMILLE were carefully tailored to better suit studies of SLA, and many of them are not simply retrieved by standard parsing methods.

On top of the data from SGATe, we used a linear regression and later an analysis of peaks of occurrence to determine the behavior of grammatical structures in the corpus. This presented us with some expected results, such as passive voice being more used in higher levels, but also showed some interesting results, like the predominance of use of connectives in C1 Level, as opposed to lower levels, as is described in the English Grammar Profile.

The new layer of annotation presented in SGATe allows for teachers to observe how learners tend to employ the grammatical content that is learned, but also allows researchers to observe how the different structures are distributed in the corpus. Considering that EFCAMDAT comprises 137 different nationalities, one further point of interest would be to

observe which type of influence the different mother tongues may have on the written production of learners of English.

Acknowledgements
The authors would like to thank the Walloon Region (Projects BEWARE n. 1510637 and 1610378) for support, and Altissia International for research collaboration.

References

Alderson, J. C. (2007). The CEFR and the need for more research. The Modern Language Journal, 91(4), 659-663.

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013, October). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project.

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. The Modern Language Journal, 91(4), 645-655.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations (pp. 55-60).

Verhelst, N., Van Avermaet, P., Takala, S., Figueras, N., & North, B. (2009). Common European Framework of Reference for Languages: learning, teaching, assessment. Cambridge University Press.

Zilio, L., & Fairon, C. (2017, July). Adaptive system for language learning. In Advanced Learning Technologies (ICALT), 2017 IEEE 17th International Conference on (pp. 47-49). IEEE.

Zilio, L., Wilkens, R., & Fairon, C. (2017a). Using NLP for enhancing second language acquisition. Proceedings of Recent Advances in Natural Language Processing (RANLP 2017) in Varna, Bulgaria, 839-846.

Zilio, L., Wilkens, R., & Fairon, C. (2017b). Enhancing grammatical structures in web-based texts. CALL in a climate of change: adapting to turbulent global conditions–short papers from EUROCALL 2017, 345.

Zilio, L., Wilkens, R., & Fairon, C. (2018). An SLA Corpus Annotated with Pedagogically Relevant Grammatical Structures. Proceedings of Language Resources and Evaluation Conference (LREC 2018) in Miyazaki, Japan.