An SLA Corpus Annotated with Pedagogically Relevant Grammatical Structures

Leonardo Zilio, Rodrigo Wilkens and Cédrick Fairon
Centre de Traitement Automatique du Langage - CENTAL
Université catholique de Louvain - UCL
{leonardo.zilio, rodrigo.wilkens, cedrick.fairon}@uclouvain.be

Abstract

The evaluation of a language learner's proficiency in second language 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 (CEFR), which addresses language learning in general and is broadly used in the European Union, while serving as reference in countries outside the EU as well. 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, SGATe, presents a rich source of information for teachers that wish to compare the production of their students with those of already certified language learners.

Keywords: SLA, syntactic annotation, CEFR, EFCAMDAT, learner profile

1. Introduction

The evaluation of a language learner's capacity when producing texts in a foreign language is not an easy task. The factors that impact the overall categorization of the produced text are many, roughly ranging from vocabulary to discourse strategies, passing by syntax and semantics. One way of facilitating this task is to have a profile of the language learner 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 the order in which 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 be evaluated as having achieved a certain level of proficiency, our objective is to look directly at the production of learners that have already been evaluated as having 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 frequency along the different language mastery levels.

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). The resource, which we named *SLA in Grammatically Annotated Texts* (SGATe), contains more fine-grained information than it would be possible to retrieve from common parsing methods, and it provides teachers with an interesting tool for comparing the written production of their own students with the annotations that are present in the corpus, which show the use of grammatical structures by certified learners.[1]

This paper is organized as follows: Section 2. describes language learning frameworks and systems that provide annotation of grammatical information; Section 3. describes the EF-Cambridge Open Language Database (EFCAMDAT) (Geertzen et al., 2013) and the annotation process, while also describing a precision evaluation of annotated structures; Section 4. presents the evaluation of a sample of the annotated data; in Section 5. we discuss the results of the annotation, by presenting more detailed information on the distribution of grammatical structures in the corpus; and Section 6. is where we present our final remarks on this study.

2. Related Work

There are different frameworks that describe how a second language should be learned and that focus on evaluating when a given learner has achieved a certain level of proficiency, such as the Common European Framework of Reference (CEFR), the Cambridge ESOL, the TOEFL, and the IELTS. For this study, the CEFR is of special relevance.

[1]The annotations were added on top of EF-Cambridge Open Language Database (EFCAMDAT) (Geertzen et al., 2013) and can be found at the EFCAMDAT Website: `https://corpus.mml.cam.ac.uk/efcamdat2/`. Alternatively, the same annotations, together with other material related to this paper, can be found at the following Website: `http://cental.uclouvain.be/resources/smalla_smille/sgate/`.

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 the teaching of foreign languages in the European Union. It provides a description of communication goals that a language 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 have the FLAIR and the SMILLE systems, both of which annotate grammatical structures based on the CEFR and use similar methods for annotating them.

The *FLAIR* system (Chinkina et al., 2016; Chinkina and Meurers, 2016) is described as an online information retrieval system that uses efficient algorithms to retrieve, annotate and rerank Web documents based on the grammatical constructions they contain. FLAIR searches online documents based on keywords selected by the user, parses the first twenty documents retrieved by the search engine and ranks them according to the settings the user selected as most important. It can recognize 87 different types of grammatical structures described in the official English language curriculum of schools in Baden-Württemberg (Chinkina et al., 2016).

The SMILLE system (Zilio and Fairon, 2017; Zilio et al., 2017a; Zilio et al., 2017b) has as its main focus the recognition of grammatical structures in online texts chosen by language learners, so that these structures can be highlighted in the text, thus aiding the learner to notice them while reading the text. SMILLE's grammatical annotation tool can recognize up to 107 different grammatical structures that were derived from Altissia's[2] pedagogical curriculum for the English language, which is based on the CEFR.

Both FLAIR and SMILLE use the Stanford Parser (Manning et al., 2014) for lemmatization, part-of-speech tagging, and dependency parsing, and then retrieve more complex, pedagogically relevant grammatical information by means of a set of rules specific to the structures that are to be recognized in the text. The main difference between the systems in terms of annotation of the grammatical information is the selection of structures that are annotated. In this regard, SMILLE presents the possibility of annotating more fine-grained structures, such as different types of gerunds and of infinitives with "to".

3. 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 et al., 2013), which presents a collection of texts produced by learners of English from different levels of proficiency. The corpus is divided according to the Common European Framework of Reference for languages (CEFR) (Verhelst et al., 2009) and contains a total of 532 thousand documents (33 million tokens) written by 83,385 learners from 137 countries. Each document 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"). The data is distributed into three main levels, each one with two sublevels (all referenced by a letter and a number): basic (breakthrough or A1, and waystage or A2), independent (threshold or B1, and vantage or B2), and proficient (effective operational proficiency or C1, and mastery or C2).

We used the SMILLE system (Zilio and Fairon, 2017; Zilio et al., 2017a; Zilio et al., 2017b) to annotate the corpus with 107 grammatical structures that are pedagogically relevant and analyzed their distribution on the different language levels. Here is an abbreviated list of the structures that the system can recognize: articles, adjectives, adverbs, personal pronouns, demonstrative pronouns, emphatic pronouns, possessive pronouns, reflexive pronouns, quantifiers, nouns, numerals, plural with special endings, plural nouns, irregular verb forms, modals, semi-auxiliaries, prepositional verbs, different types of infinitives with "to" and infinitives without "to", ellipsed infinitive, gerunds, subjects of gerund, participles, verb tenses (including perfect and continuous aspects), imperatives, passive voice, conditionals and forms of expressing hypothesis, connectives, relative clauses, sentences with "have got", question tags, short answers, wh-questions, short forms, genitives, and verb "wish" followed by past or past perfect[3].

These grammatically rich annotations on top of 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 performed an evaluation in terms of precision, which is described in the following subsection.

3.1. Evaluation of the Automatic Annotation

The annotations in the SGATe (SLA in Grammatically Annotated Texts) resource were manually evaluated in terms of precision by one linguist. This evaluation 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 in which of the annotated structures we can rely on for analyzing the annotated data. The evaluation was carried out on a random sample of the corpus: we extracted a sample

[2] www.altissia.com.

[3] In section 3.1., we present example sentences for some specific structures that we evaluated in this study.

of 40 documents from each of the 6 CEFR levels, totaling 240 documents.

Since it would be impossible to evaluate every grammatical structure that were annotated by SMILLE's system in SGATe, and since some of them simply rely on parser morphosyntactic annotation, we selected 33 structures to be evaluated that do not rely on morphosyntactic annotation[4] and that present a general overview of the features that SMILLE can annotate. Here is a list of these structures with a simple example sentence for each of them (the main words associated with the structure are marked in *italic*):

1. **Gerunds after preposition:**
   They were accused *of breaking* into a shop.
2. **Gerunds as complement of a verb:**
   We *enjoyed meeting* your friends.
3. **Gerunds instead of infinitive (no change of meaning):**
   They continued *working* hard.
4. **Gerunds instead of infinitive (change of meaning):**
   I remember *visiting* this place before.
5. **Gerunds as subject of a verb:**
   Traveling broadens the mind.
6. **Adjective + infinitive with "to":**
   I'm very *pleased to meet* you.
7. **Noun + infinitive with "to":**
   I've some *work to finish*.
8. **Verb + infinitive with "to":**
   He *refused to come*.
9. **Verbs "let" or "make" + infinitive without "to":**
   The film *made* me *cry*.
10. **Expression "let's" + infinitive without "to":**
    Let's play tennis this afternoon.
11. **Infinitive without "to" after "rather" or "better":**
    I'd *rather have* told him myself.
12. **Present perfect continuous:**
    She *has been waiting* for one hour.
13. **Past perfect continuous:**
    I *had been waiting* for one hour when the bus arrived.
14. **Future perfect:**
    I*'ll have finished* work by 5 o'clock tonight.
15. **Imperative:**
    Shut that door!
16. **Passive voice:**
    He *was seen* in London.
17. **Adverbs with passive voice:**
    The incident was *quickly* forgotten.
18. **Connectives:**
    I will call you *as soon as* I need help.[5]
19. **Relative clauses:**
    The man *who is sitting there* is my boss.
20. **First conditional:**
    If it *rains*, I'll stay at home.
21. **Second conditional:**
    If I *stopped* smoking, I could run faster.
22. **Third conditional:**
    If you *had taken* the exam, you might have passed it.
23. **Hypothesis with "would":**
    If I had more money, I *would buy* some new clothes.
24. **Hypothesis with "would have":**
    If I had studied hard, I *would have passed* the exam.
25. **Prepositional verbs:**
    I *agree with* you.
26. **Phrasal verbs:**
    Please *come in*, the doctor is expecting you.
27. **Verb "wish" followed by past:**
    I *wish* I *had* a car.
28. **Genitive marker:**
    It is *John's* book.
29. **Quantifiers:**
    There are *some* books left.
30. **Special forms of plural:**
    There are two *knives* on the kitchen table.
31. **Semi-auxiliaries:**
    We *haven't got to* read that book.
32. **Question tags:**
    It's cold today, *isn't it*?
33. **Wh-questions:**
    Why have you come so late?

4. Results

We excluded from the precision 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 annotated sample data. The overall precision of the system for the evaluated structures was 90.10% (weighed precision: 92.46%), with median at 97.50%. When we looked at the differences from level to level, we see a very bad overall precision at level A1, and then no palpable difference between the other levels, but a

[4]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.

[5]This is an example of connective of time, but several types of connectives were evaluated. Here is the full list of types of connectives: time, comparison, alternative, reason, purpose, condition, opinion, addition, explanation, and summary. Although we grouped them as one type of structure, we evaluated them also separately, as it will be further discussed in Section 4.

Table 1: Precision scores for the evaluated structures

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

much higher overall precision score, as can be seem in Table 1. 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 very good precision overall, some structures had bad performance, like gerunds as subject of verb, that had an overall precision of 55.56%, and did not perform well in any level. Other structures, like imperatives, had a not so high 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 performed very badly in Levels A1 through B1, which pulled its overall performance down to 67.53%, but actually got a nice score in the higher levels (90.20%). We verified a similar result regarding connectives, but this time in terms of granularity. The evaluation of connectives showed a 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 good performance of the other classes. The high level of precision for most of the annotated structures ensures that the annotated resource can be used for a deeper analysis of its content.

5. Profile of Grammatical Structures per Level

The SGATe (SLA in Grammatically Annotated Texts) resource comprises the entire EF-Cambridge Open Language Database (EFCAMDAT) (Geertzen et al., 2013) annotated with 107 pedagogically relevant grammatical structures. However, for diving deeper into the grammatical structures, we did not use the whole corpus, as we explain in this section. For a first exploratory analysis we used a linear regression algorithm for detecting the tendency of progression in the structures distribution along the different 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 80%.

As we observed in Section 3.1., 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 80%[6], because texts with lower scores may present some errors that can badly 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[7]. 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:

Table 2: Precision scores for the 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%

[6]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.

[7]All documents with scores above 80% were included in the C1 data. C2 documents were too few to include.

increasing tendency (angle of the line above 30 degrees), decreasing 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.

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.

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.

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), probably presenting more prominent uses in different levels, we also looked at the peaks of use of each grammatical structure. For doing this, we used the same data that was normalized by sentence, and we looked 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, present simple of the verb "to be", 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, with 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 on 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.

6. 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 a resource that we called 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.

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.

7. Acknowledgements

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

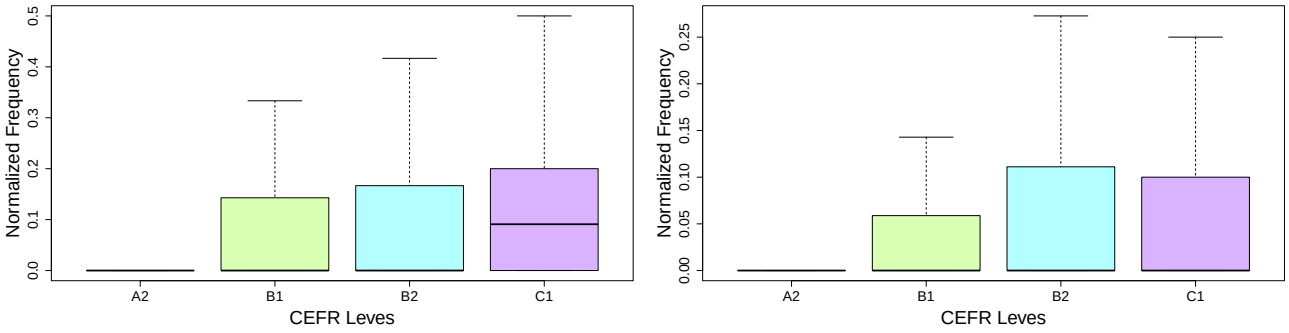
8. Bibliographical References

Alderson, J. C. (2007). The cefr and the need for more research. *The Modern Language Journal*, 91(4):659–663.

Chinkina, M. and Meurers, D. (2016). Linguistically aware information retrieval: providing input enrichment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications, San Diego, CA*.

Chinkina, M., Kannan, M., and Meurers, D. (2016). Online information retrieval for language learning. *ACL 2016*, page 7.

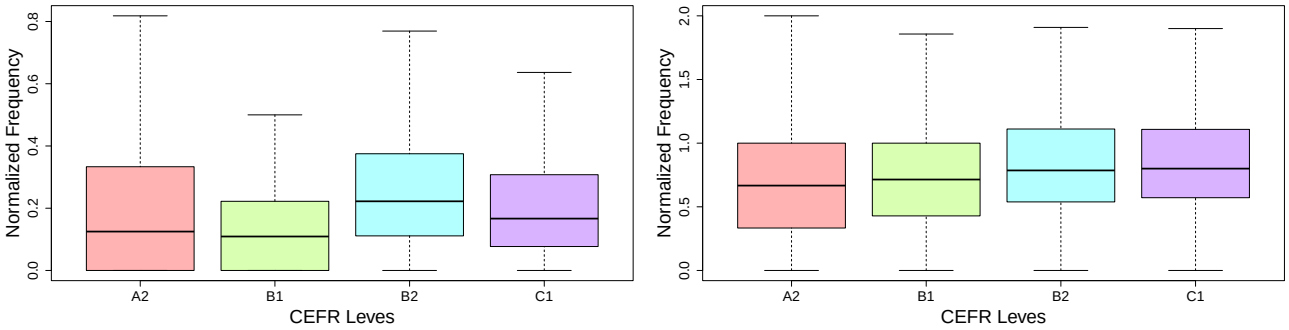
Geertzen, J., Alexopoulou, T., and Korhonen, A. (2013). Automatic linguistic annotation of large scale l2 databases: The ef-cambridge open language database (efcamdat). In *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project*.



(a) Passive

(b) Relative Clauses

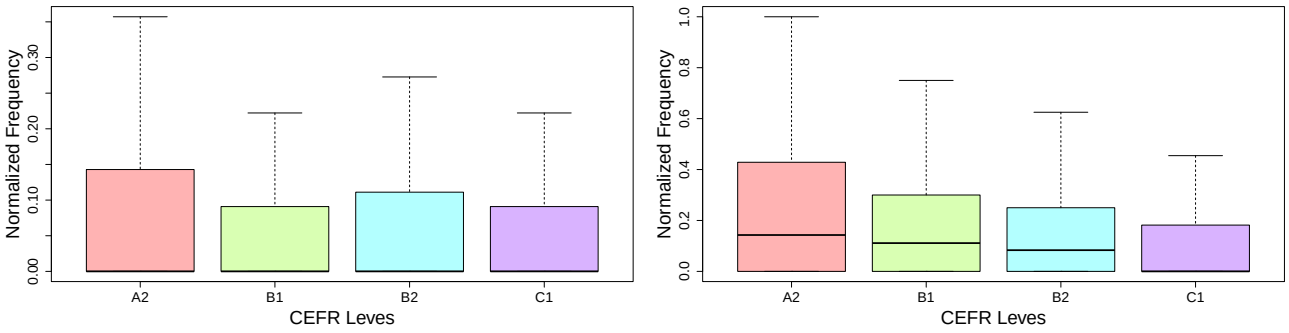
Figure 1: Examples of structures that have a tendency to be progressively more prominent along the language levels.



(a) Present Participle

(b) Present Simple

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



(a) Present Continuous

(b) Short Form

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

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. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

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

Zilio, L. and Fairon, C. (2017). Adaptive system for language learning. In *Advanced Learning Technologies (ICALT), 2017 IEEE 17th International Conference on*, pages 47–49. IEEE.

Zilio, L., Wilkens, R., and Fairon, C. (2017a). Enhancing

grammatical structures in web-based texts. In *Proceedings of the 25th EUROCALL*, pages 839–846. Accepted.

Zilio, L., Wilkens, R., and Fairon, C. (2017b). Using nlp for enhancing second language acquisition. In *Proceedings of Recent Advances in Natural Language Processing*, pages 839–846.

9. Language Resource References

Geertzen, J., Alexopoulou, T., and Korhonen, A. (2013). 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*.