Automatic extraction of English Grammar Profile constructs from the EFCAMDAT corpus

Daniela Verratti Souto
Department of Linguistics, University of Tübingen
daniela.verratti-souto@student.uni-tuebingen.de

Abstract

The Common European Framework of Reference for Languages (CEFR) has become a fundamental framework for assessing language proficiency, defining six levels of competence from A1 to C2. The English Grammar Profile (EGP) identifies grammar constructs aligned with CEFR levels. While previous work, such as that by O'Keeffe and Mark (2017), has manually analyzed Cambridge Learner Corpus data, such an approach is error-prone and time-intensive. This paper presents a pilot study utilizing the Polke extractor to automatically analyze the presence of EGP constructs in learner texts from the EFCAMDAT corpus. Three measures are used: construct presence, frequency per thousand words, and average frequency per thousand words per student. The results show the feasibility of automated analysis and suggests possible directions for further research, including validation of the levels assigned to the constructs and potential differences in levels of acquisition for speakers of different languages.

1 Introduction

The Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) has become ubiquitous in the field of language proficiency assessment, both across and outside of Europe (North, 2008). It defines six levels of language competence: A1 (Breakthrough), A2 (Waystage), B1 (Threshold) and B2 (Vantage); C1 (Effective Operational Proficiency) and C2 (Mastery). Each of these levels is characterized in terms of general functional descriptors outlining situations and tasks that a learner should be able to perform at each stage of their language development.

These descriptors are formulated as "can-do statements" and, due to the international and language-independent nature of the CEFR, do not make reference to specific grammatical or lexical items used to carry out these functions. Instead, it is up to national teams of language and teaching experts to provide reference language descriptors (RLDs) with language-specific constructs and vocabulary to complement the descriptors designed by the Council of Europe.

With this goal in mind, Cambridge University Press and Cambridge English Language Assessment, along with other institutions, initiated the English Profile Programme (EPP)[1] as an RLD project for English (see Saville and Hawkey, 2010). Since then, much work has been done to characterize learners' proficiency in English at each level in terms of grammar and vocabulary.

The English Grammar Profile (EGP) is a subproject of the EPP that aims to identify the grammatical skills of learners at each CEFR level. O'Keeffe and Mark (2017) used the Cambridge Learner Corpus (CLC) (Nicholls, 2003) to observe learner production across all levels.[2]

Although O'Keeffe and Mark (2017) base their research on empirical data obtained from Cambridge English examinations, their approach relies on a fair share of manual and qualitative work, which is error-prone and can pose difficulties when analyzing large-scale data. An automatic analysis of the level assignment for each construct is yet to be done. Such work would go a long way in confirming or adapting the construct levels based on robust methods. Similarly, by minimizing the need for human labor, a wider variety of studies can be carried out on various corpora.

The current paper is a pilot study looking at a broad range of learner texts obtained from the EFCAMDAT corpus (Geertzen et al., 2013) with the intention of observing the presence of O'Keeffe and Mark's (2017) EGP constructs at the suggested levels. To this end, I have extracted the construct occurrences in the texts using the Polke extractor,

[1] https://www.englishprofile.org/

[2] The EGP can-do statements and related information can be retrieved from https://www.englishprofile.org/english-grammar-profile/egp-online

an NLP tool currently being developed at the Hector Research Institute of Education Sciences and Psychology at the University of Tübingen. The tool has been designed for identifying the use of EGP constructs in learner texts. Once the texts were processed by Polke, I looked at construct frequency in three different ways: by their presence in texts of a given level, by their frequency per thousand words in a level, and by their average frequency per thousand words by student at a level.

2 The English Grammar Profile (EGP)

The EGP is, along with the English Vocabulary Profile, part of the EPP. In their paper, O'Keeffe and Mark (2017) describe the iterative method they employed for proposing the assignments of levels to constructs that comprise the EGP. They take the 'ELT canon of grammatical structures' as their starting point. They define this as the grammatical items that are standard in English grammar syllabi; these are familiar to teachers and are present in materials, despite often not overlapping with the categories traditionally defined by linguists. Using these categories makes sense in that the EGP was devised as a pedagogically-oriented resource and not a tool for purely linguistic inquiry.

2.1 The Cambridge Learner Corpus (CLC)

At the time of publication of O'Keeffe and Mark's (2017) paper, the CLC contained over 55.5 million words obtained from texts produced by learners taking the writing section of Cambridge English exams. The corpus contains metadata about the exam candidates (first language, gender, level of education, etc.), year of exam, exam performance and task information. Additionally, the scripts are error-tagged, which allowed the researchers to calculate the rate of correct uses of each construct.

2.2 Iterative process for assignment of construct levels

O'Keeffe and Mark describe the iterative process by which they assign a level to each construct. A summary of the steps is provided below (for further details, refer to the original paper):

1. query the corpus for a grammatical item from the ELT canon and identify different uses
2. inspect the results and check that they meet their criteria for considering the item as "acquired" at the level: frequency of use, rate of correct uses, range of users, spread of L1 families, spread of contexts of use and spread of task contexts
3. if all criteria are met, formulate a can-do statement; otherwise, move to step 4
4. search for further uses of the grammatical form until there are no more

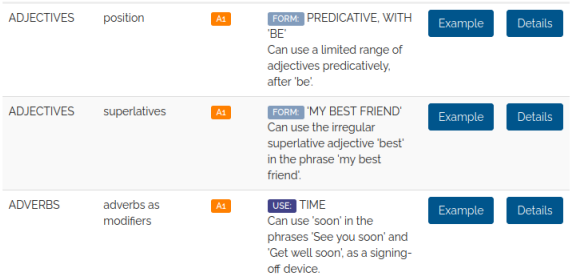


Figure 1: Example EGP descriptors for the A1 level showing the division in categories and subcategories

Ultimately, O'Keeffe and Mark derived over 1,200 empirically-motivated can-do statements, each of which is mapped to the level at which it should be acquired, and provided an example sentence for each one. This inventory of grammatical RLDs is organized in 19 super-categories with up to ten sub-categories. Figure 1 shows some examples from the EGP online website.

The process described above is painstaking and time-consuming, as well as partly reliant on manual analysis. While qualitative work like this is necessary, it is very costly and therefore difficult to replicate. Using more quantitative methods for EGP research would supplement the existing results, as well as allow for future robust and reliable analyses in different directions. For example, Hawkins and Buttery (2010) posit the existence of L1-specific criterial features. Since the EGP constructs should be criterial of the level they appear in (and higher), it would be sensible to assume that the mapping between EGP constructs and their levels could also be L1-specific. This could inform the design of learning materials and curricula for speakers of a given L1.

3 The data

The data for this study was taken from the EF-Cambridge Open Language Database (EFCAMDAT)[3] (Geertzen et al., 2013). The corpus, compiled by the University of Cambridge in collaboration with EF Education First, comprises writing

[3]Access to the corpus can be requested under the URL https://ef-lab.mmll.cam.ac.uk/EFCAMDAT.html

assignments completed by language learners enrolled in *Englishtown*, EF's online English school (which has since changed its name to *EF English Live*).

The full trajectory in *Englishtown* covers a total of 16 levels, aligned with the CEFR levels and other common proficiency benchmarks, as shown in Table 1. Every level contains eight units, each culminating in a writing task the students are required to complete. Students can qualify for a level by either being assigned to it after taking a placement test, or by completing all units of the previous level. When a student is assigned to a level, they start from the first unit and work their way up.

The version of EFCAMDAT used for this study contains over a million texts written by almost 175,000 students (Alexopoulou et al., 2017). However, this pilot analysis is based on a limited number of texts. A full analysis of the corpus with the latest version of the extractor is underway. The number of texts and words per levels from the scripts used for this study is shown in Table 2.

As can be seen by the sample of the corpus taken for the study, one of the main weaknesses of the EFCAMDAT dataset is the imbalance in terms of nationalities (and therefore L1) and of levels. A great majority of *Englishtown* students are Brazilian, which results in Portuguese as an L1 being overrepresented in the data.

The EFCAMDAT contains metainformation about learners (ID and nationality) and about the task (a general description of the task, the grade assigned to the script, the level and unit). Although not completely accurate, I have used the most widely spoken language in the country of nationality as a proxy for learners' L1. Taking this into consideration, the spread of texts across L1s is reported in Table 3

An important difference between the CLC and EFCAMDAT is that the former contains data from scripts taken from high-stakes examinations for which learners prepare by memorizing vocabulary and grammatical items deemed to be relevant for achieving a passing grade, which results in such structures being overrepresented. This kind of impact of proficiency testing on language teaching is referred to as "washback" (see Shohamy et al., 1996).

In their study, O'Keeffe and Mark (2017) adopt a particular strategy for addressing the issue of overrepresentation of features caused by exam 'display.' To determine the acquisition level of grammatical constructs, they establish a benchmark using the BNC to gauge the typical frequency of these constructs. A construct is considered "acquired" at a particular level only if its usage frequency surpasses the threshold observed in the reference corpus.

4 The extractor: Polke

Intelligent tutoring systems (ITS) and other ICALL technologies rely on representations of the learner's knowledge for the purpose of adaptivity. These L2 knowledge representations (L2KR) are used to gauge a learner's proficiency in the L2 to estimate what they have already acquired and to determine the next steps to support their language development.

A commonly used proxy for learner knowledge has been linguistic complexity (e.g. Ortega, 2003). According to Ellis (2003), complexity is defined as a measure variedness and elaborateness of some linguistic output. Although countless complexity measures have been developed and often proven to be indicative of a learner's proficiency, they are not pedagogically actionable; that is, they offer insights into language development that are too abstract for teachers to effectively use in guiding students towards improvement.

On the other hand, criterial features (Hawkins and Buttery, 2010) (like EGP constructs) are grounded in pedagogically-oriented items that have been identified by teachers and SLA researchers. Additionally, by definition, they are features that allow us to directly discriminate between competence levels and, unlike complexity measures, are easily interpreted by teachers. By extracting EGP constructs automatically, one could obtain empirically-grounded and pedagogically-informed L2KRs. This is one of the aims of the Pedagogically Oriented Language Knowledge Extraction (Polke) project[4], currently being carried out at the University of Tübingen.

Automatic criterial feature extraction is a complex issue owing to the large number of constructs that need to be identified and to the challenges that working with natural language entails, such as distinct grammatical constructs with identical surface forms. To deal with such complexities, previous studies (Meurers et al., 2010; Quixal et al., 2021) have made use of the Unstructured Information MAnagement framework (UIMA) (Ferrucci and

[4] http://www.kibi.group/project/polke

Englishtown level	1-3	4-6	7-9	10-12	13-15	16
CEFR	A1	A2	B1	B2	C1	C2
Cambridge ESOL	-	KET	PET	FCE	CAE	CPE

Table 1: *Englishtown* levels and their CEFR and Cambridge ESOL counterparts

Level	Number of texts	Number of words
A1	20413	826730
A2	13579	954170
B1	5841	541975
B2	2883	366245
C1	618	102705
C2	76	12027
Total	43410	2803852

Table 2: Number of analyzed scripts and words per CEFR level

L1	Percentage of texts
Portuguese	36.67
Chinese	30.91
Spanish	7.92
German	7.10
Italian	3.00
Russian	2.35
French	2.30
Japanese	2.03
Arabic	1.99
Korean	0.73
Others	5.0

Table 3: Percentages of analyzed texts produced by learners with different L1s

Lally, 2004) for NLP annotation, and Rule-based Text Annotation (RUTA) rules (Kluegl et al., 2016) for identifying the EGP constructs. A visual representation of the framework is provided in Figure 2.

RUTA rules make it possible to access the output from the upstream NLP annotations at a word, phrase, clause and sentence level to design grammar rules that match the EGP constructs. Studies and reports using Polke are still underway, but the technical framework draws heavily on Quixal et al.'s (2021) work. Currently, rules have been implemented for recognizing around 650 EGP constructs and a validation study is expected to be carried out in the upcoming months.

Polke's output is provided in the JSON format as a list of constructs that have been found. Each list item contains the construct ID, and the beginning and ending indices that span the text where the construct was identified. An example can be found in Figure 3. A test version of the tool is available at `http://polke.kibi.group/`.

5 Methods and analyses

5.1 Using Polke

The analysis of the texts required minimal preprocessing, as the entire NLP pipeline is integrated within the UIMA framework in the Polke server. The extractor was run on raw, uncorrected learner texts. The only assumption under which the extractor operates is that the input text consists of complete sentences or phrases. It was important not to split a sentence into separate HTTP requests, as doing so may lead to inaccurate dependencies identified by the parser. Additionally, it can only work with a limited number of words at a time. This meant that longer texts needed to be split into sections. I implemented a Python script to do the following:

1. count the words in each text file
2. if the word count is greater than 200 words, split it into sections that are as large as possible and end with sentence final punctuation
3. make an HTTP POST request to the Polke server with the appropriately-sized text sections
4. retrieve the construct ID and span from the output and store the information in a CSV, such that there is one output file per text
5. deal with errors or timeouts that may arise in the process of sending the request or obtaining the response
6. keep track of the sections that have been successfully processed

5.2 Analyzing the output

The analysis was carried out using the programming language R. The purpose was to observe, in general terms, whether the automatic extraction of constructs using Polke reflects the construct-level mappings derived in a more qualitative manner by

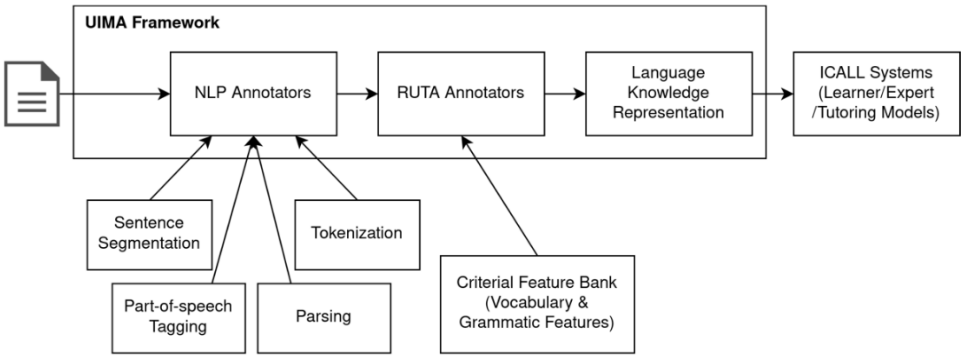


Figure 2: Obtaining L2 knowledge representations with the UIMA framework


```
{
    "annotationList": [
        {
            "begin": 0,
            "end": 19,
            "constructID": 3,
            "annotationType": "EGPConstruct"
        },
        {
            "begin": 23,
            "end": 33,
            "constructID": 37,
            "annotationType": "EGPConstruct"
        },
        {
            "begin": 23,
            "end": 33,
            "constructID": 63,
            "annotationType": "EGPConstruct"
        }
    ],
    "message": ""
}
```


Figure 3: Polke output for the text "small demonstration of the output"

expert teachers. Furthermore, by applying Polke to a large number of texts and visualizing the results in various ways, it was possible to detect irregularities in the functioning of the extractor and to report them for improvement in the latest version. Features that presented a clearly anomalous behavior and were deemed to be faulty have been removed from the analysis.

More specifically, I organized features based on their EGP level and assessed their occurrence frequency within texts at each level using three methods. These approaches are expanded on in further subsections.

5.2.1 K-means clustering of constructs

Given the methods listed above, each construct can be represented as a point in a six-dimensional space, where each dimension represents a CEFR level and the coordinates denote the frequency of the construct at that level.

Due to the large number of constructs at every CEFR level, I made use of k-means clustering for a clearer visualization of the general trends in the frequency of use across levels. K-means clustering is a clustering algorithm whose aim is to divide M points in N dimensions into K clusters so that the Euclidean distance between points in the same cluster is minimized (Hartigan and Wong, 1979). For our purposes, M refers to the number of constructs belonging to a CEFR level and N is equal to 6.

A shortcoming of the k-means clustering algorithm is that the number of clusters to be found, K, needs to be determined by the analyst. I employed the elbow method to obtain sensible values for K (see Cui, 2020 for an in-depth explanation of the algorithm). The method consists of calculating the variance within clusters, termed Within-Cluster Sum of Squares (WCSS), with several values of K. In a better clustering, the WCSS is lower. By plotting the WCSS against the different K values, it is possible to observe a "bend" (or "elbow") where the decrease in WCSS becomes less dramatic, as seen in Figure 4. This is generally regarded as the optimal number of clusters. In the cases when the position of the elbow was unclear, I chose the highest of the candidate options as the K value.

Using a line graph, I plotted the centroids of the clusters to find general trends in the output. The term "centroid" refers to the mean of all points in a cluster; in other words, centroid are the center of mass, or "average", of each cluster. The plots (in Figure 9) show that the trend is for the use of constructs to increase along with the levels.

After applying the clustering to each metric, I inspected the features in each cluster (see Section 6.2). There were no clear commonalities in the contents of the can-do statements within the same

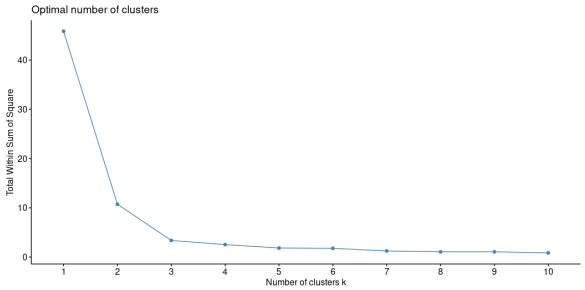


Figure 4: An elbow plot with the elbow at 3

cluster other than their similar frequency of use. A visual representation of the clusters can be found in Figure 8. Plots for A2 and B2 are shown as means of illustrating what was done.

5.2.2 Percentage of presence in texts

The first approach I explored consisted in analyzing the extractor output and observing the proportion of texts at each level that contain each construct. In other words, for each construct-text pair there were two options: either the construct was used in the text or it was not.

For visualizing the behavior of all constructs, I used line plots like the ones shown in Appendix A. Each line represents an EGP construct, the x-axis contains the text levels, while the y-axis represents percentage of presence of the feature in texts of a level.

Taking a binary approach like the one described here, where the actual frequency in the documents is not relevant, can be advantageous as it avoids dealing with complex issues regarding normalization. However, it deviates considerably from the approach taken by O'Keeffe and Mark, which heavily relies on the normalized frequencies of constructs. As a result, I repeated the analysis to deal with frequencies per thousand words.

5.2.3 Frequency-based approaches: per level and per student

There are several ways in which normalized construct frequencies can be calculated. Firstly, I opted for treating all texts in a level as if they were one. This has the advantage of circumventing issues related to task effects (Alexopoulou et al., 2017), since scripts from all tasks in the level were taken into account. Then, calculating the frequency per thousand words of each construct was trivial. Example plots can be found in the appendix.

While for the purpose of observing the appearance of each feature across texts of all six levels this approach yields satisfactory results, it would be useful to have information about variance in future research. As a result, I implemented an algorithm to calculate the frequency with which learners at each level use every construct. The algorithm is described below.

1. at each level, find all students that have completed a minimum of n texts[5]
2. calculate the frequency with which each student used each construct
3. plot the mean frequency of each construct

Since for any given construct we have as many frequencies as we have students, measures of spread like variance and standard deviation can be calculated by future researchers.

6 Results

6.1 Analyzing all features

I visualized all features assigned to each CEFR level by plotting their presence in texts at all levels according to the three different metrics. Figure 6 in the appendix shows the results for A2 and B2.

An interesting observation is that, for all three analyses, A2 features seem to reach higher frequencies than B2 features (note the differences in scales for the y-axis). This seems a sensible outcome, as learners at A2 are taught basic constructs that are fundamental for expressing both simple and complex ideas, whereas, generally speaking, more advanced learners acquire less frequent grammatical items whose function is to better express nuance and are of use mostly when discussing topics with higher complexity or in more formal registers. In other words, constructs taught at basic levels like A1 and A2 are needed all across the six CEFR levels since they are frequent in all language production, whereas grammatical elements acquired later are often less frequent in general usage. Similarly, comparing 6a and 6b, it is apparent that A2 features are generally consistently present from low to more advanced levels, whereas B2 features are less prominent until roughly the B1 mark.

Interestingly, the features in Figure 6a as well as in 6b seem to increase their frequency from the onset until C1, and then drop at C2. It is possible that

[5]For lower levels, n = 24, the total of tasks at the CEFR level, was good enough for getting a fairly large number of texts. However, at higher levels, fewer students had completed all tasks, so it n had to be lowered considerably.

this is due to beginner and intermediate learners becoming more comfortable with the constructs they have acquired at prior stages and using them more often as their proficiency increases; however, once they approach mastery of the language, they prioritize a varied repertoire of grammatical means of expression, leading to a decrease in the frequency of less advanced items. Nevertheless, this peak at C1 (globally or with respect to the surrounding levels) is also present in the plots for some C2 features, as shown in Figure 7. Manual inspection of the data and the Polke output would be required for a satisfactory explanation for this pattern.

6.2 K-means clustering of features

Figure 8 shows a visualization of the constructs as points in a vector space clustered by their distance as described in Section 5.2.1. The graphs show that, for features of the same level, the clustering looks somewhat similar even though frequencies are calculated using different summary statistics. For instance, 8a (obtained from the percentage of presence analysis applied to A2 constructs), 8c (A2 constructs, per level) and 8e (A2 constructs, per student) show clear similarities in the lower x-axis despite the different number of total clusters suggested by the elbow method. The similarities are even clearer in the B2 plots in Figures 8b, 8d and 8f, with features quite consistently remaining in the same cluster across different analyses, like 740 and 932 forming a cluster, as well as 859, 899 and 935. It is worth noting that, while 8f seems to differ considerably from the other two B2 clusterings, closer inspection shows that the differences are due to the visualization tool reducing the six-dimensional space into two dimensions differently, and a rotation of the graph brings out the similarities with the other two analyses.

These similarities across the clusterings using different summary statistics suggest that the three approaches yield somewhat similar results when it comes to comparing constructs to one another in terms of their frequency, so the ultimate choice of a metric for further analysis is likely to hinge on factors other than the clusters that emerge from the use of one summary statistic or another.

6.2.1 Looking inside the clusters

So as to obtain a general overview of what kind of information the clusters capture, it is of interest to inspect the grammatical constructs that are considered similar by the k-means clustering algorithm. By way of example, I summarize the clusters found for B2 EGP constructs obtained by pooling together all texts in a level (as seen in Figure 8d). Below is a discussion of the clusters, both in terms of how their frequencies develop across text levels and of their linguistic characteristics.

Cluster 1 Cluster 1 (marked in red in Figures 8d and 9d) contains a total of 47 constructs, making it by far the largest cluster for this level but also the most dense in the vector space.[6] As can be seen in Table 4, the criterial features in this cluster are distributed across 15 of the 19 possible EGP super-categories. This wide variety in grammatical elements is a strong indicator that the constructs at hand might not share many linguistic similarities beyond the frequency with which they appear. A closer look at the individual EGP constructs that constitute this cluster confirms that they are, in fact, very diverse, ranging from simple negation of some auxiliaries (construct 801 [7]), to complex constructions such as passivization of ditransitive verbs using the indirect object in subject position (construct 873) or the combination of modal verbs with perfective or progressive aspects (constructs 836,[8] 849,[9] 977[10]).

To address the development of these features across the six CEFR levels, the centroids of the cluster have been plotted in Figure 9d. They are represented by the red line. Although in comparison to the other centroids it appears flat, Figure 5 shows it in isolation, revealing that it is practically absent until A2, then increases dramatically until C1, where it peaks at around 0.055, and then drops to roughly 0.026 at C2. It is worth noting that the centroid does not represent any particular construct; instead, it denotes the point in the vector space that minimizes the distance between itself and the constructs within the cluster.

Cluster 2 Cluster 2 contains only two constructs: 740, "[c]an use adjective phrases to modify nouns," and 932, "[c]an use 'one' as a generic personal pronoun in the subject position to mean people in general." These two grammatical elements differ

[6] The high density in this area of the vector space makes it difficult to further separate these constructs into more clusters, as increasing the value of k breaks the 'looser' clusters into individual constructs and keeps the 47 points in question together.

[7] Can use the negative form with 'will'

[8] Can use 'must have' + '-ed'.

[9] Can use 'should be' + '-ing'.

[10] Can use 'could' have + '-ed'.

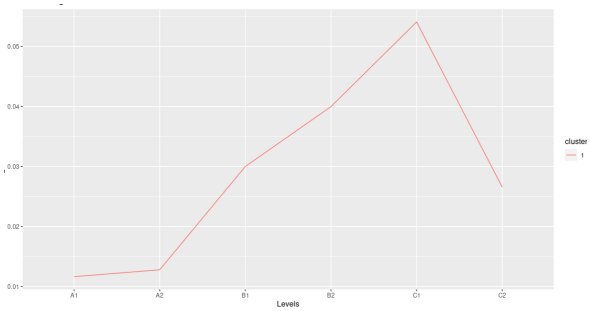


Figure 5: Zooming in on the centroid for cluster 1

considerably from each other in terms of their linguistic features, with the first one referring to the formation of more elaborate syntactical structures for noun modification (adjective phrases instead of simple adjectives), whereas the latter corresponds to the use of a generic pronoun. The fact that both these constructs are in the same cluster is surprising; however, Figure 8d shows that they are quite far apart from each other in the vector space, suggesting that they ultimately share little in common and, with a higher value for k during the k-means clustering, would have likely been assigned to distinct clusters.

The centroid of the second cluster (green line in Figure 9d) shows that this cluster is by far the most frequent. It differs from the other centroids in that its frequency at A1 is significantly higher than at A2. It is construct 932 that is responsible for this shape, as it reaches 3 occurrences per thousand words at A1 and then drops to roughly 1.5 at A2. While pinpointing the exact reason for this irregular behavior is impossible without manually inspecting the A1 scripts, I would hypothesize that this is due to task effect (Alexopoulou et al., 2017) for texts of a certain unit, or a result of students copying (or "lifting") part of a text that was provided for carrying out a particular task. It would then make sense to encounter a local peak at A1 and then observe the usual increasing curve as the levels of the texts advance.

Cluster 3 Cluster 3 contains three features, each belonging to a different EGP super-category. Construct 859 refers to using the '-ing' form of the verb in subject position; 899 corresponds to using the past simple after certain subordinating conjunctions like 'if', 'since' and 'while', and construct 935 pertains to the usage of some indefinite pronouns (like 'each', 'either' and 'several') as subjects and objects. Notably, two out of these three EGP constructs revolve around noun phrases headed by nonnouns (verbs in the former and pronouns in the latter). Their general development across text levels can be seen from the blue centroid plotted in 9d. It shows that their frequency steadily increases until reaching its peak at B1, only to drop at more advanced levels, likely due to learners acquiring alternative ways of expressing ideas and avoiding overuse of more basic grammatical elements.

An interesting observation is that using the method of pooling all texts in each level results in a frequency peak at B1, that is, before the level assigned to the EGP; however, the same does not hold for the percentage of presence in texts, as show in Figure 9b. In any case, both analyses reveal that this cluster of features is more prevalent at B1 than at B2, which might indicate that some adjustments could be made to the level assignment of these EGP constructs.

Cluster 4 There are six constructs in this cluster:

- using 'be' + 'to' in the present to refer to the future (construct 791)
- using the auxiliary 'will' in the affirmative (construct 797)[11]
- using a pronoun or noun + 'be' + adjective + 'to' + (past) infinitive (construct 807)
- using the present perfect passive affirmative / negative form, often for reporting (constructs 869 and 870)
- using the passive with a range of tenses and ditransitive verbs with the direct object in subject position and the indirect object in a prepositional phrase

The passive features prominently in the cluster, as well as two different ways to refer to the future. This makes it the most cohesive cluster in terms of the linguistic characteristics of the members contained in it. As the purple line in Figure 9d shows, it observes a relatively steady increasing pattern of acquisition until its peak at C1, then slightly decreasing at C2.

All in all, the clustering seems to only be sensitive to the frequency of the constructs and only in rare cases do they reflect similarities in terms of the linguistic attributes of the criterial features that compose each cluster.

[11] Note that the negative form of 'will' is also a B2 construct (801), but it is in cluster 1.

EGP Super-category	Clusters			
	1	**2**	**3**	**4**
Adjectives	4	1	-	-
Clauses	3	-	-	-
Conjunctions	1	-	-	-
Determiners	2	-	-	-
Future	3	-	-	2
Modality	9	-	-	1
Negation	1	-	-	-
Nouns	3	-	1	-
Passives	7	-	-	3
Past	3	-	1	-
Present	1	-	-	-
Pronouns	4	1	1	-
Questions	1	-	-	-
Reported speech	1	-	-	-
Verbs	4	-	-	-
Total	47	2	3	6

Table 4: Count of B2 constructs per EGP super-category present in each cluster using frequency per thousand words at each level

7 Conclusion and outlook

This paper explores the idea of automatically extracting EGP constructs from learner corpora such as EFCAMDAT, a corpus annotated with levels that can be aligned with the six CEFR levels, using the Polke extractor. The data visualization shows that, in general terms, EGP constructs of lower levels are present across all scripts, regardless of the level of the text. However, for criterial features belonging to higher levels, an increase in frequency can be seen for texts around the proficiency level assigned to the feature by EGP. This suggests that using Polke for quantitatively analyzing the mapping between constructs and levels proposed by the English Grammar Profile is viable and can be pursued further.

Future work in this area could focus on quantitatively validating the level assignment of individual constructs by using the Polke extractor on different corpora annotated by proficiency level. Additionally, since the analysis could be done automatically, a variety of options could be pursued, such as defining L1-specific mappings between constructs and levels. Such work would allow the design of educational materials and tools that are better adapted for learners of different backgrounds.

Acknowledgements

I would like to extend my gratitude to the individuals who have contributed to the completion of this term paper:

Firstly, I am grateful to Nelly Sagirov for providing me with starter code that facilitated the process of sending requests to Polke.

I would like to express my gratitude to Prof. Detmar Meurers for his guidance throughout this project. His assistance in selecting the topic and valuable insights into data analysis methodologies were greatly beneficial.

Additionally, I wish to acknowledge Dr. Xiaobin Chen for his constructive feedback, which played a significant role in refining the content. Furthermore, I am thankful to Dr. Chen for sharing Figure 2, which allowed to better illustrate the functioning of the extractor.

References

Theodora Alexopoulou, Marije Michel, Akira Murakami, and Detmar Meurers. 2017. Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, 67(S1):180–208.

Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Mengyao Cui. 2020. Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance*, 1(1):5–8.

Rod Ellis. 2003. *Task-based language learning and teaching*. Oxford university press.

David Ferrucci and Adam Lally. 2004. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10(3-4):327–348.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 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*, pages 240–254. Citeseer.

John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108.

John A Hawkins and Paula Buttery. 2010. Criterial features in learner corpora: Theory and illustrations. *English Profile Journal*, 1:e5.

Peter Kluegl, Martin Toepfer, Philip-Daniel Beck, Georg Fette, and Frank Puppe. 2016. Uima ruta: Rapid development of rule-based information extraction applications. *Natural Language Engineering*, 22:1–40.

Detmar Meurers, Ramon Ziai, Luiz Amaral, Adriane Boyd, Aleksandar Dimitrov, Vanessa Metcalf, and Niels Ott. 2010. Enhancing authentic web pages for language learners. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 10–18.

Diane Nicholls. 2003. The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In *Proceedings of the Corpus Linguistics 2003 conference*, volume 16, pages 572–581. Cambridge University Press Cambridge.

Brian North. 2008. The educational and social impact of the CEFR in Europe and beyond: a preliminary overview. In *Language testing matters: investigating the wider social and educational impact of assessment-proceedings of the ALTE Cambridge Conference*, pages 357–378.

Lourdes Ortega. 2003. Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing. *Applied Linguistics*, 24(4):492–518.

Anne O'Keeffe and Geraldine Mark. 2017. The English Grammar Profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics*, 22(4):457–489.

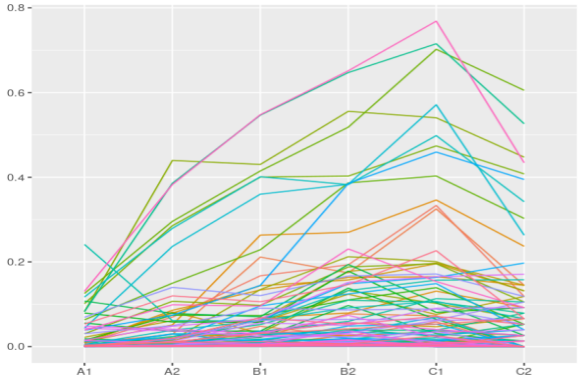
Martí Quixal, Björn Rudzewitz, Elizabeth Bear, and Detmar Meurers. 2021. Automatic annotation of curricular language targets to enrich activity models and support both pedagogy and adaptive systems. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 15–27.

Nick Saville and Roger Hawkey. 2010. The English Profile Programme–the first three years. *English Profile Journal*, 1:e7.

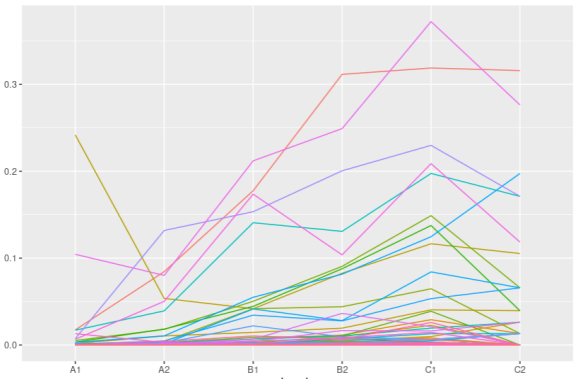
Elana Shohamy, Smadar Donitsa-Schmidt, and Irit Ferman. 1996. Test impact revisited: washback effect over time. *Language Testing*, 13(3):298–317.

A Appendix

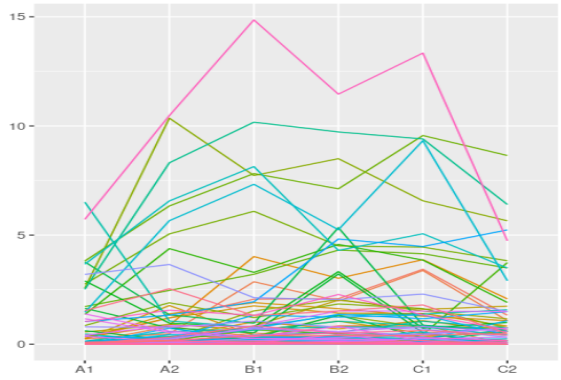
This appendix contains the plots for construct visualization mentioned in the main text.



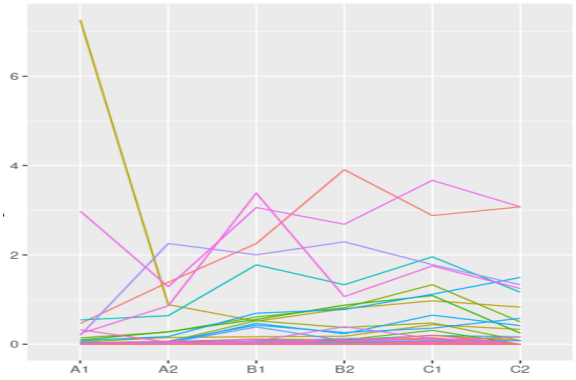
(a) Line plot for A2 features: percentage of presence



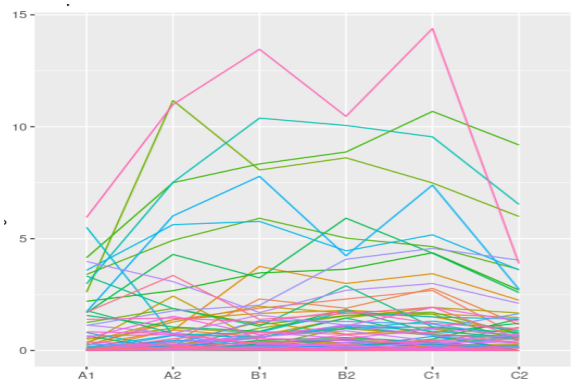
(b) Line plot for B2 features: percentage of presence



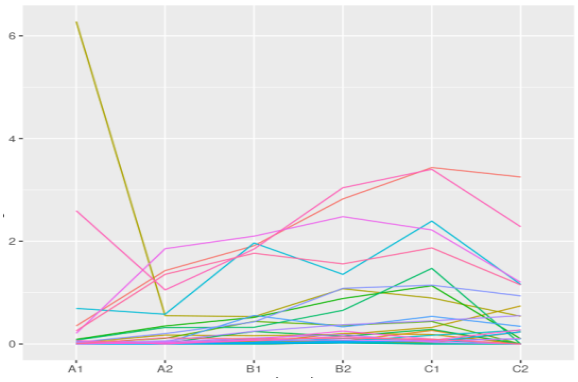
(c) Line plot for A2 features: per level



(d) Line plot for B2 features: per level

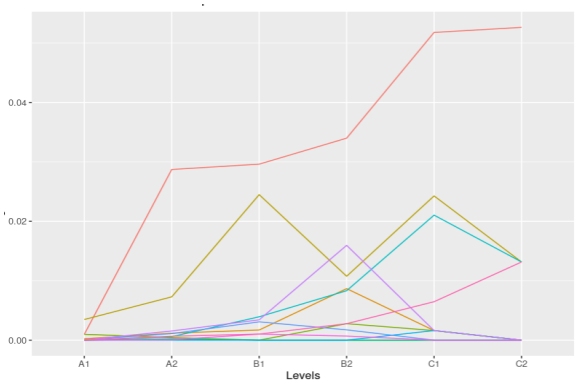


(e) Line plot for A2 features: per learner

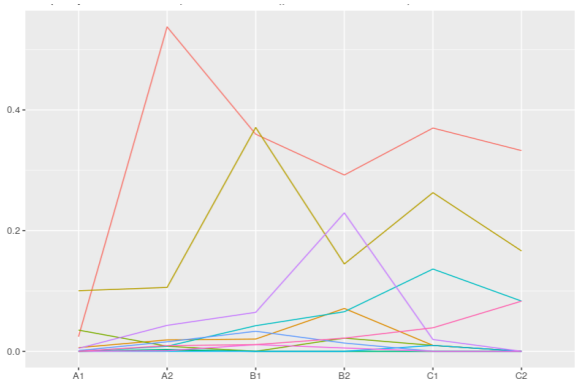


(f) Line plot for B2 features: per learner

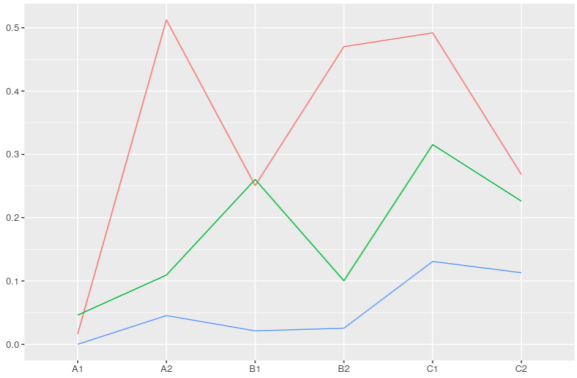
Figure 6: Plots showing the frequency of all constructs per text level



(a) Line plot for all C2 features, calculated by percentage of presence

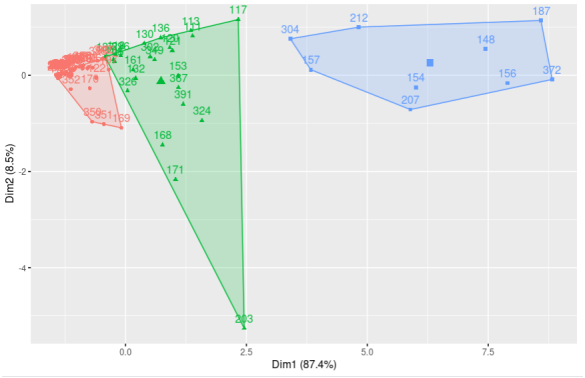


(b) Line plot for all C2 features and their frequencies per a thousand words, calculated per aggregated texts at each level

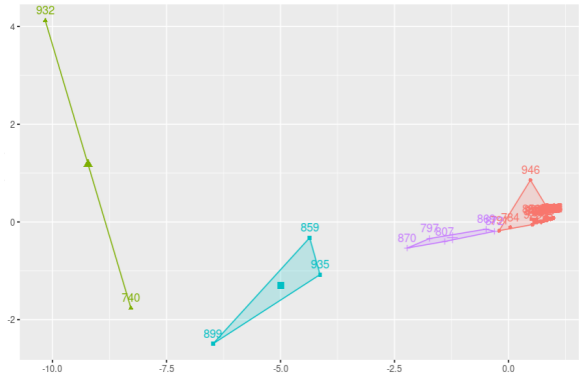


(c) Line plot for all C2 features and their average frequency per student at each level.

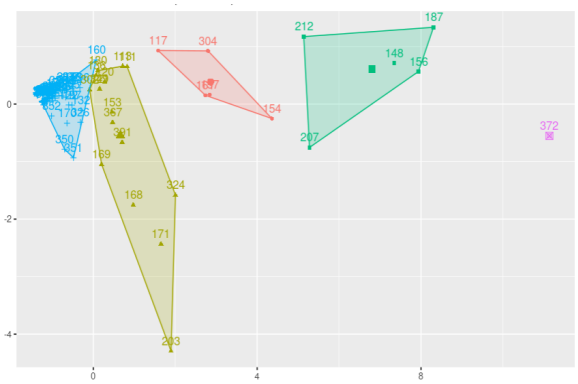
Figure 7: All C2 features analyzed by percentage of presence and per level



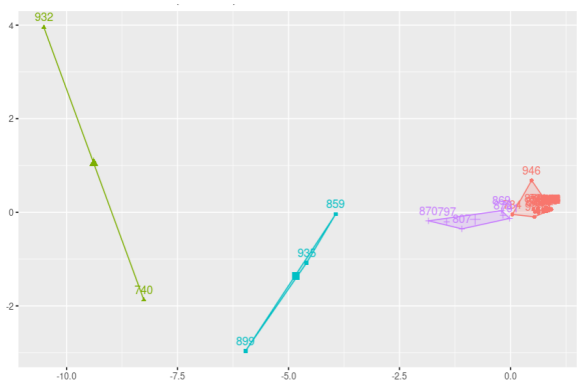
(a) Clustering of A2 features: percentage of presence



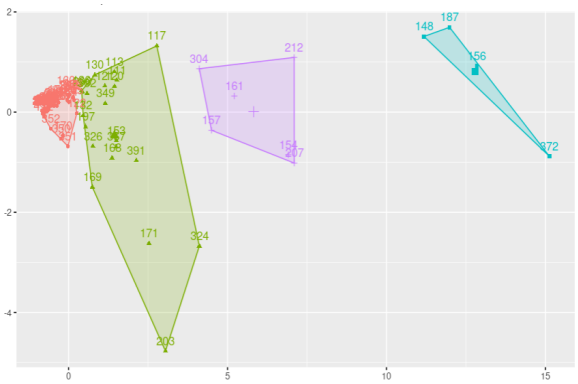
(b) Clustering of B2 features: percentage of presence



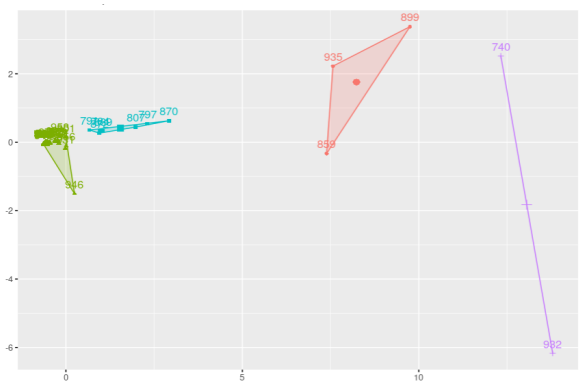
(c) Clustering of A2 features: per level



(d) Clustering of B2 features: per level

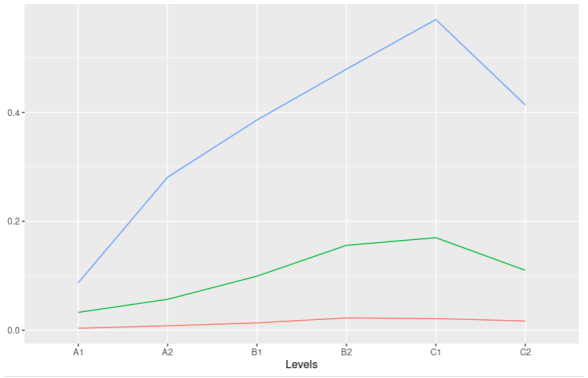


(e) Clustering of A2 features: per student

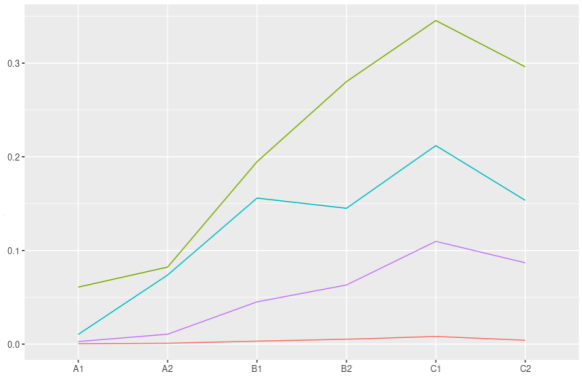


(f) Clustering of B2 features: per student

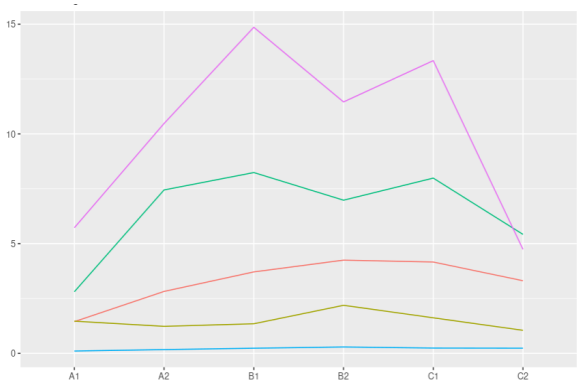
Figure 8: K-means clustering of constructs



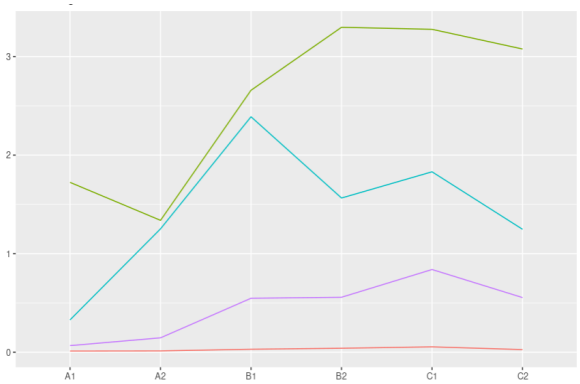
(a) Centroids of A2 features clusters: percentage of presence



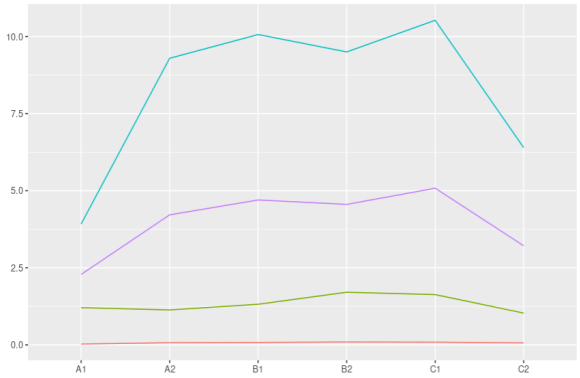
(b) Centroids of B2 features clusters: percentage of presence



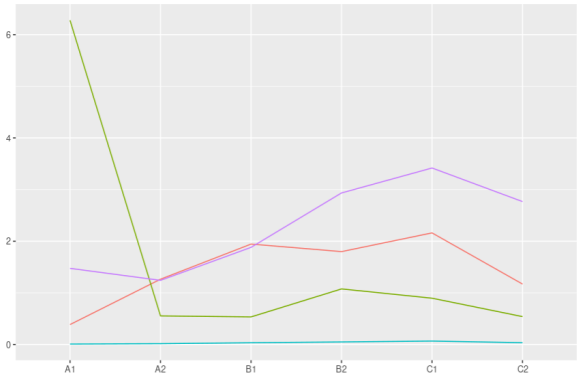
(c) Centroids of A2 features clusters: per level



(d) Centroids of B2 features clusters: per level



(e) Centroids of A2 features clusters: per student



(f) Centroids of B2 features clusters: per student

Figure 9: Line plots showing the centroid of A2 and B2 constructs with different methods