# Language learning analytics : designing and testing new functional complexity measures in L2 writings

**Thomas Gaillat**
LIDILE, Université de Rennes / Rennes, France
thomas.gaillat@univ-rennes2.fr

## Abstract

This paper presents the initial stage in the design of an ICALL system. The objective is to develop a system that automatically generates linguistic analytics of L2 learner writings. Student texts will be processed with NLP tools producing different types of textual measures. We present the design of a new functional complexity metric aiming to capture the paradigmatic competition between forms mapped to the same communicative function, i.e. microsystems. More precisely, we analyze the variations of the FOR and TO prepositions in terms of frequency and probability of occurrence. Relative frequency shows significant correlations with CEFR levels suggesting its possible use in an analytics report system. More work is required to extend the approach to other microsystems.

## 1 Introduction

When using an L2, learners make assumptions about form-function mappings. They observe contexts in order to understand the meanings of specific forms. "The task facing the learner is to discover (1) which forms are used to realize which functions in the L2 and (2) what weights to attach to the use of individual forms in the performance of specific functions." (Ellis, 1994, p.375). In completing this task, learners modify their internal L2 system, gradually stabilise the mappings and improve proficiency.

Proficiency has been the focus of much research and it relies partly on the use of the *complexity* construct. Grammar complexity features form a major part of the elements used to operationalise this construct. Two ways of operationalising the construct have emerged. One based on holistic measures factoring in several grammar constituents such as the ratio between the number of dependent clauses and the total number of clauses in a text. The other one relies on frequency counts of different grammar patterns classified in terms of complexity. For all the benefits in both approaches, neither operationalises the variations between multiple forms mapped to one function. Previous work suggests that there are variations in mappings across proficiency levels (O'Keeffe and Mark, 2017). So capturing these mapping variations could help to identify factors of proficiency in L2 learners.

Form-function mappings could be operationalised as probability indicators in the use of one form over other forms mapped to the same function. These indicators could be generated by models stored in the expert module of an Intelligent Tutoring System (ITS). To achieve this, the models must be built with data trained on occurrences of the forms. In this paper, we present an illustration of the design of a new functional complexity measure operationalising the FOR, TO prepositions mapped to the communication function of "expressing purpose". We design a measure generated with a probabilistic model which is intended to be part of a proficiency predictor system.

## 2 Theoretical background

Structure complexity is a construct that includes functional complexity as one of its sub-types (Bulté and Housen, 2012, p.25). This construct relies on the mappings between forms and functions of linguistic forms. It has been operationalised in various ways such as specific parts of speech or dependency relations (Settles et al., 2018) or syntactic constituents as in CTAP's feature selector module (Chen and Meurers, 2016). The use of functional complexity features offers two advantages for studies in the field of Second Language Acquisition. First, based on learner corpora, these features can be used to design metrics exploited for modelling purposes in prediction tasks such as CEFR classification (Vajjala, 2018; Kyle, 2016;


Thomas Gaillat. Language learning analytics : designing and testing new functional complexity measures in L2 writings. *Proceedings of the 11th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2022)*. Linköping Electronic Conference Proceedings 190: 55–60.

Pilán et al., 2016; Yannakoudakis et al., 2011). Secondly, they can be exploited for the design of specific linguistic feedback which is meaningful for learners and teachers (Riemenschneider et al., 2021).

Learners make confusions between forms of the same communicative function. They tend to hesitate between one form or the other when they want to express a specific function such as obligation, probability, purpose or reference. These hesitations illustrate one aspect of the competition model in which learners constantly resolve conflicts while choosing forms (MacWhinney et al., 1984), hence the notion of L2 microsystems. These microsystems are unstable as learners unexpectedly group forms that do not necessarily fall in the same functional paradigm (Py, 1980). Due to this instability in the mappings, the microsystems are transitional in nature (Gentilhomme, 1980). They include erroneous mappings which later are removed, leading the learner to better proficiency.

The microsystems can be analysed according to their paradigmatic relations. The following examples show the use of the FOR and TO prepositions in contexts expressing purpose, and more precisely followed by verbs and nominalised verbs with ING. In all three cases, the learners present difficulties to choose the right preposition within to-clause or prepositional phrase contexts. In (1), a more acceptable choice might have been FOR. There seems to be a confusion between the use of the complement to-clause controlled by *write* and the use of a prepositional phrase (PP) introduced by FOR. In (2) and (3), the learner clearly misused FOR instead of TO. In (2) the learner shows a confusion between the use of a complement extraposed to-clause and a PP. In (3), the confusion seems to be between a PP and an adverbial clause wrongly introduced by FOR (instead of *in order to* for instance).

1. Dear Mr or Madam : I am writing *to* enquiring about the possibility of requesting a loan (Sentence ID: 41038:1 Teaching level: 10 Learner nationality: Spain)

2. But, sincerly, I think that it's a strategy *for* promote his new movie. (Sentence ID: 3762:2 Teaching level: 7 Learner nationality: Spain)

3. Then, you go to the sport centre *for* doing sport. After, you walk the dog and you give it the food . (Sentence ID: 16950:7 Teaching level: 6 Learner nationality: Spain)

The underlying assumption in these examples is that there is an L2 specific microsystem in which FOR and TO compete paradigmatically to express purpose, be it in to-clauses or prepositional phrases including ING noun phrases. In the context of L2 automatic analysis, a challenge is to quantify the variations within this microsystem and others, which leads us to the following research questions: How can we capture variations between forms mapped to the same communicative function? Which form variations can be observed within a microsystem across CEFR levels? Answering these questions with computer models would provide the ground for the design of an NLP pipeline.

## 3 A learner language analytics system

The microsystem approach falls within a broader objective, i.e. the design of an ICALL system (see Figure 1) for teachers. The objective is to develop a computer system that automatically generates linguistic analytics of learner writings. The students will input their texts which will be processed with NLP tools producing different types of textual measures, some of which microsystem based. The system will provide visualisations of the measures for teachers to analyse their students' writing profiles.

Developing the system requires the validation of the textual measures in terms of correlations. A method to identify correlations between linguistic features and metadata including proficiency, task types, learning habits will be applied. This paper discusses the case of the statistical validation of the FOR, TO microsystem.

## 4 Method for the validation of the measures used in the system

### 4.1 Data

We used the Spanish subset of the EFCAMDAT corpus (Geertzen et al., 2013). It is made up of 8,187 texts written by EnglishTown students based in Spain. Table 1 provides the breakdown. The data was annotated in terms of 16 proficiency levels which can be converted in the six CEFR levels as described in the corpus manual[1].

[1] Available at https://corpus.mml.cam.ac.uk/faq/EFCamDat-Introrelease2.pdf (last access 24/11/2022)

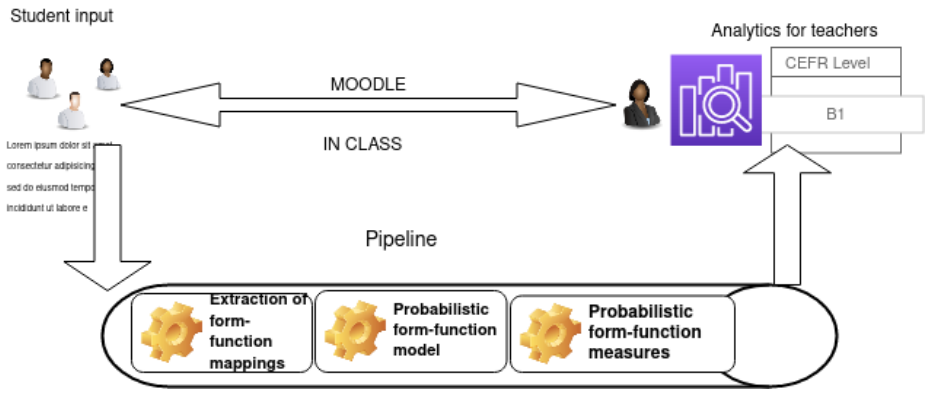


Figure 1: NLP pipeline - from data collection to visualisation

| CEFR level | Writings | Mean of words |
|---|---|---|
| A1 | 2,571 | 106.75 |
| A2 | 2,065 | 91.41 |
| B1 | 2,004 | 120.3 |
| B2 | 1,175 | 174.1 |
| C1 | 340 | 193.8 |
| C2 | 32 | 195 |

Table 1: Number of writings and mean number of words/text across CEFR levels in the Spanish subset of the EFCAMDAT learner corpus

### 4.2 Pre-processing and extraction

The texts were pre-processed with UDPipe (Straka et al., 2016) using the Stanford *english-ewt-ud-2.5-191206* English model in R. The tool provides grammatical annotation such as PoS, lemma, dependency relations and morphological features linked to the class of words (gender, number, case...). The CEFR levels were then appended to the resulting dataset.

The objective of the extraction was then to identify TO and FOR prepositions related to the function of "purpose". To extract the forms we proceeded twofold. Firstly, we only focused on actions (nominalised with ING or not) and retrieved verbs of any tense or aspect following the two forms. Secondly, following (Biber et al., 1999, p.693-751) on the identification of complement to-clauses, we applied queries that identified the forms according to a predetermined list of verbs and adjectives controlling to-clauses. We filtered by semantic class (Biber et al., 1999, p.700-705) keeping speech act verbs (e.g. ask, tell, warn), verbs of desire (e.g. hope, wish, like), verbs of intention or decision (e.g.decide, choose); verbs of effort (e.g. try, manage, fail). In the case of adjective controlling to-clauses, we filtered those referring to willingness (Biber et al., 1999, p.718). For the identification of prepositional phrases introduced by FOR and adverbial to-clauses (introduced by *in order to*, *so as to* or *to*, the heuristic identified forms immediately following a noun (plural or singular).

To measure extraction performance, we randomly sampled 100 occurrences of each form from dataset resulting from the first step. Each of these forms was then manually tagged as a purpose-related form or not. We then applied the heuristic to automatically identify the purpose-related forms. We then computed Recall, Precision and F1 metrics as shown in Table 2.

| Forms | Precision | Recall | F1-Score |
|---|---|---|---|
| TO | 0.56 | 0.42 | 0.48 |
| FOR | 0.73 | 0.66 | 0.69 |

Table 2: Precision, Recall and F1-Score for the extraction of FO and TO related to the purpose function

After the first step, we extracted 497 occurrences of FOR and 13,772 occurrences of TO. Applying the aforementioned heuristic resulted in a dataset of 9,820 occurrences of FOR (N=300) and TO (N=9,520). The distribution of the forms across levels is presented in Table 3.

### 4.3 Testing the significance of relative frequencies of microsystems as potential features of proficiency

To test the validity of the microsystem as a construct varying with proficiency, we analysed the relative frequencies of occurrence of the two forms (per 1,000 words) across the CEFR levels. We computed a one-way ANOVA to verify whether

| CEFR | TO | FOR |
|---|---|---|
| A1 | 581 | 14 |
| A2 | 1,328 | 43 |
| B1 | 2,934 | 116 |
| B2 | 2,441 | 91 |
| C1 | 849 | 28 |
| C2 | 69 | 0 |

Table 3: Distribution of FOR TO prepositions across CEFR levels

the differences between groups were significant.

### 4.4 Testing the significance of probabilities of microsystems as potential features of proficiency

We also wanted to measure the impact of microsystem internal probabilities on proficiency. To this end, we built a binomial logistic regression to model the microsystem forms. As we had an imbalanced number of forms, we first randomly extracted an even number of each preposition (N = 300 * 2). This was intended to prevent the classifier from assigning too much weight to the dominant class. We then split the dataset into a training (75%) and a test set (25%). The model was built with the *multinom()* function in the *nnet* R library (Venables and Ripley, 2002). We also computed a one-way ANOVA to verify whether the differences in the means of the probabilities between CEFR groups were significant.

## 5 Results

### 5.1 Relative frequencies as features

To test the significance of relative frequencies of the microsystem forms, we analysed their variations. We computed the means of frequencies in the texts across the six levels. Figure 2 shows the results. As frequencies of FOR were very low we plotted a barchart of the means. There seems to be a distinction in the use of TO at the A1 and C1 levels compared with the other levels. The use of TO seems to gradually decrease as proficiency increases. Regarding the FOR preposition, it appears to be favoured at the B1 level compared with the other five levels.

The one-way ANOVA for the TO prepositions reveals that the differences between the means of the CEFR groups are significant (F-value = 9.7, $p < 0.001$, Adjusted $R^2 = 0.01$) with an extremely low effect size. The ANOVA for the FOR prepositions shows that differences in means are not significant across the CEFR groups (F-value = 1.09, $p > 0.05$, Adjusted $R^2 = 0$).

### 5.2 Probabilities as potential features

To obtain relative probabilities of one component over the one one, we built a binomial model with the two microsystem prepositions as dependent variables, and parent and adjacent POS as independent variables. We first tested its classification power. The predicted probabilities of the TO vs FOR preposition (reference level) were extracted and matched to the true CEFR level of each observation of the test set. The model performance indicators show a 0.97 accuracy (95% CI (0.93-0.99) and p-value $< .001$). Precision and recall were 0.97 and 0.97 respectively.

We then analysed the distribution of the probabilities of TO vs FOR across the true CEFR levels in the fitted model over the training set. Figure 3 shows the variations of the data points including their variance and medians. If the variations overlap, medians appear to be quite distinct between levels. For instance, TO seems to be more likely to occur than FOR in the A1, A2 and C1 levels. The distribution of the FOR preposition is indirectly plotted as *1-P(TO)*, where *P(TO)* stands for probability of TO, i.e. a less-that-50% probability of TO implies a more-than-50% probability of FOR.

The one-way ANOVA showed that the differences between the means in the probabilities across the six CEFR groups are not significant (F-value = 1.49, $p > 0.05$, Adjusted $R^2 = 0$). A closer analysis shows that probabilities of the B1 level show $p = 0.05$.

## 6 Discussion and future work

In this paper, we have presented a new functional complexity metric which attempts to operationalises the paradigmatic competition between the TO and FOR prepositions used in the same communicative function which is "expressing purpose". The objective is to evaluate the metric as a proficiency criterial feature. This metric could be introduced in an ICALL system dedicated to generating analytics reporting measures of communicative functions for language teachers.

The experiment included the extraction of FOR and TO used with a meaning of purpose. The re-

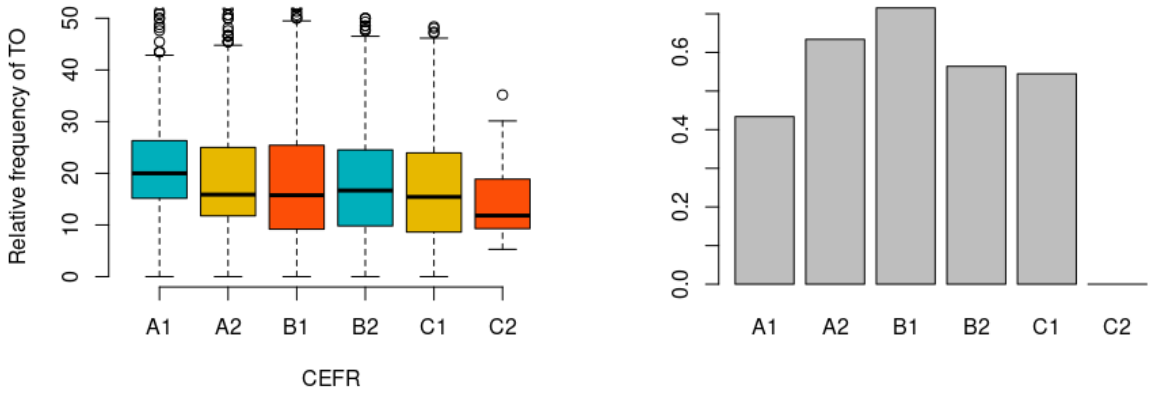


Figure 2: Distribution of relative frequencies of TO (left) and distributions in means of relative frequencies of FOR across CEFR levels in the EFCAMDAT Spanish subset

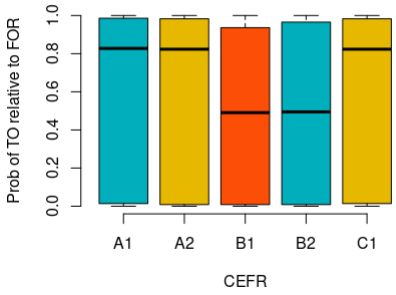


Figure 3: Fitted probabilities of the TO vs FOR prepositions across CEFR levels in the training set of the EFCAMDAT Spanish subset

sults are mixed. The extraction of FOR appeared to give good results while the extraction of TO proved to be a challenging task. In order to capture all possible learner uses (correct and incorrect), the heuristic is based on a list of words appearing in adverbial or complement to-clauses or in prepositional phrases introduced by FOR. The list needs further refinement regarding words introducing to-clauses. For instance, post-verification of the annotated sample showed a number of inconsistencies such as the presence of "have to" as a purpose expression.

The experiment's main objective was the statistical validation of the metric in terms of mean difference between the CEFR levels. The assumption was that if there were significant differences, the metric variations could be used as features of the system. We obtained mixed results. The model provides good classification power. The distributions of both the relative frequencies and the binomial logistic regression probabilities show variations across CEFR group. However, only the TO relative frequencies are significant, albeit with an extremely low effect.

These findings suggest that there are issues to solve before the metric could be used as a predictor in new texts. More testing needs to be done in order to validate the approach. Ultimately, the new measure should be tested as a feature in a proficiency predictor model. Finer-grained microsystem patterns could also be identified thanks to the work on the English Grammar Profile (O'Keeffe and Mark, 2017).

More microsystems are being designed. Following Gaillat et al. (2021), modals, articles, deictics are some of the forms that will be tested. The next stage is to create a program generating microsystem measures as part of a pipeline (see Figure 1). This pipeline will output its results in a MOODLE module (Dougiamas and Taylor, 2003) in the form of indicators linked to linguistic communicative purposes. Teachers will be able to interpret and diagnose their learners' linguistic profiles.

## References

Douglas Biber, Stig Johanson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Longman, Harlow.

Bram Bulté and Alex Housen. 2012. *Defining and Operationalising L2 Complexity*. John Benjamins Publishing Company.

Xiaobin Chen and Detmar Meurers. 2016. CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 113–119. The COLING 2016 Organizing Committee. Event-place: Osaka, Japan.

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

Rod Ellis. 1994. *The Study of Second Language Acquisition*. Oxford University Press, Oxford, United Kingdom.

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

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2013. Automatic Linguistic Annotation of Large Scale L2 Databases: The EF-Cambridge Open Language Database (EFCamDat). In *Proceedings of the 31st Second Language Research Forum*, Carnegie Mellon. Cascadilla Press.

Yves Gentilhomme. 1980. Microsystèmes et acquisition des langues. *Encrages*, (Numéro spécial):79–84.

Kristopher Kyle. 2016. *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication*. Dissertation, Georgia State University, Georgia.

Brian MacWhinney, Elizabeth Bates, and Reinhold Kliegl. 1984. Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning and Verbal Behavior*, 23(2):127–150.

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. Publisher: John Benjamins.

Ildikó Pilán, Elena Volodina, and Torsten Zesch. 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2101–2111, Osaka, Japan. The COLING 2016 Organizing Committee.

Bernard Py. 1980. Quelques réflexions sur la notion d'interlangue. *Revue Tranel (Travaux neuchâtelois de linguistique)*, 1:31–54.

Anja Riemenschneider, Zarah Weiss, Pauline Schröter, and Detmar Meurers. 2021. Linguistic complexity in teachers' assessment of German essays in high stakes testing. *Assessing Writing*, 50:100561.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Sowmya Vajjala. 2018. Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, (28):79–105. ArXiv: 1612.00729.

W. N. Venables and B. D. Ripley. 2002. *Modern Applied Statistics with S*, fourth edition. Springer, New York.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Stroudsburg, PA, USA. Association for Computational Linguistics.