Exploring a New Grammatico-functional Type of Measure as Part of a Language Learning Expert System

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

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

Abstract

This paper explores the use of L2-specific grammatical microsystems as elements of the domain knowledge of an Intelligent Computer-assisted Language Learning (ICALL) system. We report on the design of new grammatico-functional measures and their association with proficiency. We illustrate the approach with the design of the IT, THIS, THAT proform microsystem. The measures rely on the paradigmatic relations between words of the same linguistic functions. They are operationalised with one frequency-based and two probabilistic methods, i.e., the relative proportions of the forms and their likelihood of occurrence. Ordinal regression models show that the measures are significant in terms of association with CEFR levels, paving the way for their introduction in a specific proform microsystem expert model.

1 Introduction

This paper explores the use of L2-specific grammatical systems as elements of the domain knowledge of an Intelligent Computer-assisted Language Learning (ICALL) system. Such systems rely on Natural Language Processing approaches that conduct several high-end tasks such as Grammatical Error Detection (GED), automatic reformulation or proficiency level prediction. As part of the Intelligent Tutoring System (ITS) category, they rely on models that have have an expertise, which is language use in their case.

Expert models encapsulate the domain knowledge which is required to describe the learner's language skills involved in tasks such as writing production. In ITSs, there are several possible strategies used to acquire and represent domain knowledge (Nkambou et al., 2010). Among those are rule-based cognitive models, describing learning strategies, and Constraint-based models (CBM) describing principles that rely on correct solutions to a problem.

In the case of ICALL, representing the knowledge of learners has traditionally been done within the Constraint-based model (CBM) framework thanks to correct-usage principles derived from native language use. For instance, some Grammatical Error Detection tasks are processed on the basis of target hypotheses (TH) (Lüdeling and Hirschmann, 2015), i.e. the correct version of what is meant by a learner in a specific segment. In this type of tasks, correct versions of erroneous segments or patterns are compared with the TH (Bryant et al., 2017) to identify incorrect uses. As useful as it has proved to be, this type of approach tends to reduce the knowledge about L2 language production to what native speakers would say or write by focusing on error correction. In doing so, it overlooks the meta-knowledge that language learning experts possess regarding acquisition processes. Experts' evaluations of learner language not only rely on TH, but also on what they know of the grammatical, lexical, semantic and pragmatic features in L2 writings of different proficiency levels, be they negative or positive features (Bulté and Housen, 2012). This allows them to position the learner's productions in terms of level and to provide feedback.

We argue that an expert ICALL system should not be reduced to error identification and correction on the basis of native language production, but include comprehensive knowledge about the range of L2 linguistic profiles at different stages of language learning. We intend to use such profiling as part of a learning-analytics system providing in-

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 466–476

formation to teachers on their learners' linguistic developmental stages.

Modelling the domain knowledge with such profiles linked to proficiency levels is necessary. In order to do so, we draw inspiration from rule-based cognitive models. The role of rule-based cognitive models is to describe the knowledge involved in "student performance in a given task domain, including strategies, problem-solving principles, and knowledge of how to apply problem-solving principles in the context of specific problems" (Aleven, 2010). When applied to language learning, this approach complies well with describing the strategies used to elaborate language patterns, including idiosyncrasies. This makes rule-based cognitive models quite comprehensive in describing learner language characteristics.

Our proposal follows this approach, as it considers an expert model as a cognitive entity that knows positive and negative characteristics of an L2 set of writings at various stages of proficiency. The expert model should not simply "know" the rules that operate for native speakers, it should also include the probability of patterns that govern specific levels. Many grammatico-functional microsystems (MS) exist that describe a part of the grammatical reasoning at work in production. They are convenient to describe the psychological reality of the learner and may be linked to proficiency as in the English Grammar Profile (O'Keeffe and Mark, 2017).

As an illustration of the broad process, we present the design and implementation of a specific linguistic microsystem as a rule-based cognitive model, namely the THIS, THAT and IT proform microsystem. Our working hypothesis revolves around the idea that different proficiency levels prompt different linguistic contexts around the use of the microsystem, which leads to different odds of using the forms in the microsystem. Therefore, by observing the probability of using a given form in the microsystem as a function of the context, we could capture aspects of the learner's grammatical reasoning that points to a given proficiency level. This approach raises two research questions:

1. What is the likelihood of microsystem forms in L2 writings, according to the linguistic context that surrounds the microsystem?

2. What is the distribution of these probabilities across CEFR levels?

To answer these questions, we propose to use a model to describe the probabilities of use of THIS, THAT and IT proforms depending on context, as a first step toward modelling linguistic profiles. In other words, this model predicts the likelihood of a learner using either THIS, THAT or IT given the linguistic context of this proform, while being agnostic to whether the choice of such proform was correct or not. To assess the relevance of using the likelihood of microsystem forms as linguistic profiling, a second model predicts proficiency levels using only the probabilities of using THIS, THAT and IT output by the microsystem prediction model. If this second classification model can discriminate proficiency levels using only the predicted probabilities for the forms of the microsystem, then the microsystem likelihood model is a coherent way to build domain knowledge indicators for profiling. In section 2 we present the theoretical background underlying our research. Section 3 presents the data and the microsystem extraction methods used to exploit it. In section 4, we present how the microsystems are implemented and evaluated in terms of extraction and predictability. Section 5 covers the results obtained with different modelling approaches to validate the associations between microsystem and proficiency.

2 Theoretical background

ICALL systems are ITSs, and it is relevant to understand the distinctions between the types of expert models before reviewing the types of ICALL models.

Expert models in general ITSs. Intelligent Tutoting Systems require expert models which fall into three main categories, i.e. black-box, glass-box (or rule-based) and cognitive models (Nkambou et al., 2010; Anderson, 2013). Following Aleven (2010), we place rule-based models alongside CBMs as part of the cognitive category. Black box models are said to be inexplicit in their representations as they only provide the final results (Nkambou, 2010, p.18) and show correct input-output behaviour with very little use for their internal computation (Anderson, 2013, p.26).

Cognitive models show different degrees of interpretability, which is useful for instruction delivery. Their decision making processes lend themselves well to giving feedback to learners. In the subcategory of CBMs the requirements that all

solutions should satisfy are set in advance rather than having to map all possible errors and correct solutions. This simplifies the search space by narrowing down the possible solutions and avoiding breaking the domain principles. Conversely, rule-based models rely on an comprehensive set of rules that can be deterministic or probabilistic. The rules mirror the way an expert analyses a problem by taking into account positive or negative observations.

A good expert model seems to revolve around several principles. Not only does it have to produce correct results, but it also needs to have high cognitive fidelity, i.e., the compliance of its decision making features with those that are used by learners. In addition, the expert model must filter out the feature space "according to the same restrictions as a human does" (Anderson, 2013).

Expert models in ICALL. The aforementioned distinctions can be used to understand the different types of expert models that exist in ICALL systems. Are they black boxes or cognitive models? Rule-based or CBM? Depending on the tasks and the adopted NLP approaches, they may fall into one of the categories, showing or not their cognitive inclination.

As far as we know, most second language models employ supervised learning methods which rely on very different types of features. Neural approaches with text embeddings and transformer models provide very accurate results in different tasks such as Grammar Error Detection (GED) (Bryant and Briscoe, 2018) or Automatic Essay Scoring (AES) (Rama and Vajjala, 2021). However, the rules and features they rely on are difficult to interpret, turning them into black boxes and leading to poor cognitive fidelity.

A number of GED tasks have relied on supervised learning approaches based on error coded datasets including corrected statements as target hypotheses (Settles et al., 2018). These hypotheses may be seen as the CBM principle, i.e. reference points with which learner language is compared (Bryant et al., 2017). By way of edit-distance metrics, the models can be used to provide error identification in context. However, they cannot explain the reasons for the errors. Their decision making process does not rely on information that is cognitively meaningful for learners.

Other supervised-learning models are based on probabilistic rules relying on explicit linguistic features. In proficiency prediction tasks, a number of experiments were conducted with models relying on morphosyntactic and lexical features (Tack et al., 2016; Pilán and Volodina, 2018; Yannakoudakis et al., 2018). These linguistic features make up intelligible rules that have a degree of cognitive fidelity. However, in spite of their linguistic characterisation, some are not very actionable by teachers. This is due to the complexity of their design in terms of variables (Gaillat, 2022). For instance, the Automated Readability Index (ARI; (Smith and Senter, 1967)), a measure of difficulty in reading, is composed of two variables (Average Sentence Length (ASL), the Average Word Length (AWL)) whose combination in a formula[1] is hard to interpret. Because they are not designed to provide any other feedback than the result, these models do not have high cognitive fidelity.

Some advanced Automated Writing Evaluation (AWE) systems show greater cognitive fidelity as they try to match their feedback with interpretable linguistic information. Based on linguistic features used in supervised learning methods, the systems can contextualise the errors with grammatical justifications (Attali and Burstein, 2006; Yannakoudakis et al., 2018). Some Automatic Essay Scoring systems, which rely on semantic and discourse complexity metrics, feed from their expert models' features to elaborate feedback messages on cohesion for learners (Dascalu et al., 2013). In these cases, the connection between the models' rules and the wording of the feedback messages shows a focus for high cognitive fidelity. While elaborating the messages, these systems rely on expert models that filter out irrelevant knowledge that could impair the cognitive reception by learners. One important aspect is that Dascalu et al. (2013) add specificity as an extra dimension. They use two specific models for two specific tasks, i.e. a view of cohesive links in discourse and a view of stance variation in discourse.

Our proposal follows the same principle applied to the grammatical rather than the cohesive dimension. The objective is to design expert models that capture the hesitations that learner may have on specific syntactic paradigms. For instance, learners may hesitate between different determiners, or they may have confusions in the use of demonstrative pronouns. We intend to design

[1] ARI = 0.5ASL + 4.71AWL - 21.34

several expert models for microsystems specifically linked to linguistic functions. Their goal is to provide fine-grained knowledge of the variations between forms of the same function. Microsystems reflect the learners' hesitations that are part of the competition model in which learners constantly resolve conflicts while choosing forms (MacWhinney et al., 1984). These hesitations create microsystem instability as learners unexpectedly group forms that are not necessarily mapped to 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. Following Gaillat et al. (2021), we focus on the referential proform microsystem made up of THIS, THAT and IT. The purpose is to compute how each of these forms, mapped to the same referential function, is likely to occur in relation to its two other competitors.

3 Data pre-processing and proform extraction

3.1 Data

The proform microsystem measures are computed with data extracted from the the EF Cambridge Open Language Database (EFCAMDAT) corpus (Geertzen et al., 2013). This corpus results from the collaboration between the Department of Theoretical and Applied Linguistics at the University of Cambridge and Education First (EF). The data was collected on EF EnglishTown, an online school. Our data set is made up of 1,180,507 texts written by students in 191 countries around the world. The data was annotated in terms of 16 proficiency levels which were converted into the six CEFR levels as described in the corpus manual[2]. Table 1 shows the distribution of the average number of words per text and per level.

The data was pre-processed with the methods detailed in Section 3.2. Then, the Grew pattern extraction explained in Section 3.3 was applied, and only the samples where an instance of the microsystem was found are kept. This results in a table that contains 881,627 samples, i.e., as many lines as there are occurrences of proforms IT, THIS and THAT in the EFCAMDAT learner writings. This table also contains 726 columns

CEFR	Writings	Mean of tokens	SD
A1	626,005	39.32	21.46
A2	308,014	68.82	24.42
B1	168,473	98.88	30.23
B2	61,366	137.27	43.67
C1	14,709	171.13	49.03
C2	1,940	176.98	71.95

Table 1: Descriptive statistics of EFCAMDAT writings across CEFR levels

corresponding to the linguistic features about the environment of the microsystem.

3.2 Pre-processing

Prior to microsystem extraction, the data are annotated according to the Universal Dependencies (de Marneffe et al., 2021) framework. The annotations notably include Universal Dependency tagged part-of-speech, lemmas of tokens, and morphological features such as case, number, gender, etc. Linguistic annotations were obtained with the UDPipe pipeline (Straka et al., 2016) using the English model trained on the GUM corpus[3] (Zeldes, 2017). This model shows reliability for POS and dependency annotation on L1 and L2 (Kyle et al., 2022).

3.3 Proform extraction with Grew pattern queries

Grew (Amblard et al., 2022) is a graph rewriting tool that manipulates linguistic representations and is aimed at natural language processing applications. It is used to extract the elements of a microsystem from a sentence, given its linguistic annotations.

Grew creates an annotated graph from a CoNLL-U annotated sentence, with the words and their linguistic annotations (lemma, xpos, upos...) as nodes, and the dependency relations between the words as edges. Using a set of patterns, it is then possible to isolate only the words in the graph that follow these patterns. We create patterns corresponding to proform usage of IT, THIS and THAT. Example (1a) shows the THIS pattern. The heuristic searches for all tokens which are DEPendent on a GOVernor predicate by a dependency relation of the following types: nominal subject (nsubj) in a passive voice structure (:pass), oblique (obl), nominal modifier, object, conjunct, or root

[2] Available at https://corpus.mml.cam.ac.uk/faq/EFCamDat-Intro_release2.pdf (last access 25/03/2023)

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

of sentence (see de Marneffe et al., 2021, p.266-267).

(1a)

```
THIS_PRF::DEP[wordform="this"
|"these"|"This"|"These"];
GOV-[nsubj|obl|nsubj:pass|nmod
|obj|nsubj:outer|conj|root]
-> DEP.
```

As a results, proforms, such as in examples (2a) and (2b), can be extracted.

- 2a) This song may be a joke now , between musicians , but at the time *it* came , *this* rocked .
- 2b) *That* is how I found the class of Sciences of Education in Paris 2 . I went to the global opening and when I was listening to the presentation of the classes , I was sure *this* was what I wanted to study for my future .

Once the patterns have been extracted, information about the linguistic environment of the target microsystem is collected, including morphological, syntactic and part-of-speech information available for the words in a five-word window around the target proforms. The same type of information is also collected for the dependency governor of the target word, as well as the distance along the dependency tree from the target word to the root of the sentence.

Evaluation of the extractions To check the soundness of using Grew patterns to extract microsystems, we conducted the following evaluation. All the sentences that contain an occurrence of the words IT, THIS or THAT, whether they are proforms or not, are selected from the preprocessed data. For each of these words, 100 samples are selected randomly among those that contain the forms, or the maximum amount available if it is less than 100. Additionally, some samples, not containing any of the forms, are also selected. This results in 358 samples used only for evaluation of Grew patterns. This sampling strategy is different from the modelling-evaluation strategy applied in Section 4.3 because, here, it is essential to capture forms of any function. On the contratry, the modelling strategy solely requires proform samples of the forms.

The gold standard is set by an expert who annotates whether the identified form is indeed a IT, THIS or THAT proform, or none of those. The samples are also independently run through Grew. The tool outputs the patterns for the identified forms which are then compared to the expert annotations. A notable feature of this data is the unbalance of the forms, with the THIS proform making out only 2% of the annotated samples, and 60% of the samples displaying no use of the microsystem. Note that a subsequent development of this study will include three annotators with measures of inter-rater agreements.

The weighted F1-score of Grew extractions reaches 0.82, with a weighted precision of 0.90 and a weighted recall of 0.80. This shows that using Grew patterns as a tool to identify the microsystems is viable, and does not select many forms that are not indeed part of the microsystem. A word of caution is to be given about the results for each individual form: while the IT proform occurrences are almost always perfectly identified, and most THAT forms found by Grew are indeed correct, many relevant THAT proforms are not identified in the text, as shown in Figure 1. This phenomenon can be explained by the sample selection strategy : sentences that contained the string of characters THAT were selected in this data set. However, the word THAT covers a wide range of other functions than proform, namely, subordinator, relativizer, adverbial, demonstrative determiner. IT, on the other hand, is more often used in its referential function in spite of its possible other functions, i.e., impersonal use, extrapositional use, cleft use and expressing weather/time/distance (Huddleston and Pullum, 2002, p. 960, Biber et al., 1999, p. 332). The samples containing THAT are therefore less likely to contain a large proportion of proform use of THAT, contrary to the samples containing IT. On inspecting extraction errors of THAT, it also appears that the proform use is confused with the relativizer use of the form. To address this problem, the Grew extraction query of THAT proform should be revised with a finer-grained filtering strategy. Still, with a grain of salt concerning the extraction of THAT, these results show that Grew is a relevant tool for the extraction of the elements of the microsystem.

Moreover, this first evaluation also provides some insight on the rarity of proform uses of THIS, THAT and IT, highlighting variability in the frequencies of use. IT is more often used in its

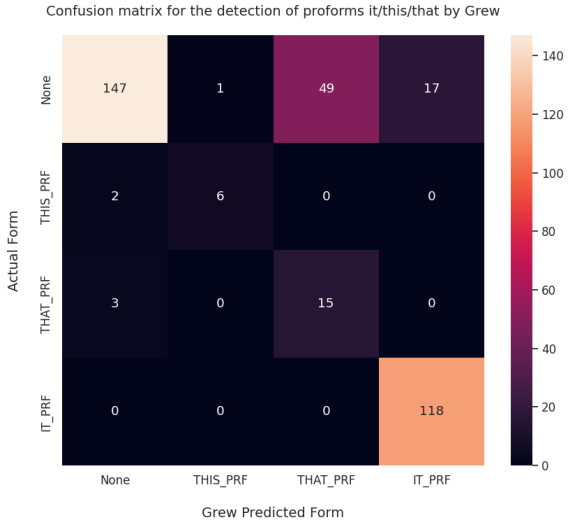

Figure 1: Confusion matrix for the evaluation of Grew pattern extraction

proform function than THAT, for instance. This draws the attention on the need to address this imbalance in crafting the statistical models, as such models are often biased by unbalanced data, and to analyze our results in the light of these uses of the proforms.

4 Design of the proform microsystem measures

4.1 Conceptual design

The conceptual idea of a microsystem is that each form is used relative to its competitor forms because they are mapped to the same referential function. For instance, one possible assumption about the proform microsystem is that the use of a THAT is detrimental to the use of IT and THIS. In order to identify the best operationalisation of the microsystem concept, we identified three types of measures capturing the forms' relative variations. The measures are based either on proportions or probabilities of occurrence.

Regarding proportions, we tally the counts of each IT, THIS and THAT for each writing, then create the percentage for each MS as in:

$\text{MS}(x_{ij}) = f_{ij} / \sum_{k=1}^{n} f_{kj}$

where x = the microsystem, i = the ith component of the microsystem, j = the jth text, n = the total number of forms in microsystem and f_{ij} = the frequency of a component in text j.

Regarding probabilities, we apply two types of models. First, a multinomial logistic regression model predicting forms on the basis of linguistic features extracted from the forms' local contexts.

$\sigma^{-1}(p(i|C)) = \alpha + \beta_1(\text{c}_1) + ... + \beta_n(\text{c}_n) + \epsilon$

where $p(i|C)$ is the probability of observing a proform i knowing the context C made up of features $C = \{c_1, ..., c_n\}$ and σ^{-1} is the logit function.

Second, a neural network predicting the probabilities of a form given the linguistic environment. Given an input sample C that represents the linguistic environment of a form, the goal is to compute the conditional probability of observing one of the forms of the microsystem, i.e, THIS, THAT or IT.

$p(i|C) = \sigma(f_2(LR(f_1(C))))$

where $p(i|C)$ is the probability of observing a proform i knowing the context C as defined above, $f_k(c) = cA_k + b_k$ are linear layers with trainable parameters A_k and b_k, $LR(c) = \max(0, c) - 0.01 \times \min(0, c)$ is a Leaky ReLU activation function, and $\sigma(c)_i = {e^{c_i}}/{\sum_{j=1}^{K} e^{c_j}}$ is the softmax activation function. The input C consists of the one-hot-encoded categorical variables in the linguistic environment of a form. The LeakyReLU activation function has been preferred over the ReLU function as a way to mediate the issues of vanishing gradient during training, induced by the sparse feature representation of the input due to one-hot-encoding.

4.2 Technical implementation

The relative proportions are based on the raw frequencies of the proforms in each text and are computed on all the texts.

In the case of the logistic regression measuring approach, the model relies on the following features: POS, Universal Dependency information regarding heads, POS of tokens found in a [-5;+5] position interval and dependency distance between a form and its head. As not all variables of the data set were assigned values (especially morphological features which are dependent on word types) variables with more than 10% missing values are dropped.

In the case of the Neural measuring approach, all available linguistic annotation is collected as features at first: POS, morphological features and Universal Dependency information of tokens found in the [-5;+5] position window and dependency distance between a form and its head. Then, only features where more than 60% of the samples are not null were kept. This is done as an

	A1	A2	B1	B2	C1	C2	Total
IT	18360	20291	18406	11482	2476	371	71386
THAT	9268	16320	25009	14950	5009	830	71386
THIS	19372	16821	20673	10565	3518	437	71386
Total	47000	53432	64088	36997	11003	1638	214158

Table 2: Number of samples for each proform at each CEFR level in the balanced training data set

attempt to reduce unnecessary dilution of the information through one-hot-encoded variables that would be mostly null. The network is trained over 50 epochs, using the Adam optimiser with a 0.0005 learning rate.

4.3 Evaluation

To evaluate the accuracy of the measures given by the proform predictive approaches, we split the data into 80% training and 20% testing, for a total of 705,302 training samples and 176,325 test samples. The training data set is then balanced with regard to the number of THIS, THAT or IT forms, in order to avoid model imbalance. We take random samples of IT and THAT equal to the number of THIS (i.e., 71,386 occurrences, the lowest number among the three proforms), resulting in 214,158 training samples. Details on the composition of the training data set are provided in Table 2.

We evaluate the three measure construction methods, that is, proportions, logistic model probabilities and neural network probabilities, in two steps. Firstly, we examine the predictive capacity of the systems used to create the measures : if these models cannot properly classify the proform given a certain context, then they are not likely to create good measures for the microsystem. We therefore perform multinomial logistic regression and train the neural network approach on the training data and predict labels in the testing data, using linguistic features listed in Section 4.2.

In a second phase, to evaluate whether the measures correlate with proficiency, we perform modelling with ordinal logistic regression as a descriptive model. Taking as descriptors the probabilities of using THIS, THAT and IT output by the previous model, we investigate whether there is an association between these measures and the odds of increasing CEFR level.

5 Results and discussion

5.1 Measure creation

We separately inspect the three approaches used to create the measures. The first proportion-based approach can only provide an insight in the tendency of the learners to use the different forms of the microsystem, as it is a count-based method and not a statistical model. The other two approaches can be evaluated with the usual accuracy, prediction and recall scores, presented in Table 3.

Beginning with measures based on proportions, Figure 2 depicts the distribution of IT, THIS, THAT relative proportions across CEFR levels. It shows a reduction in the percentage use of IT as CEFR level increases. The reverse is seen in the percentage of THAT use, with an increase at higher CEFR levels. A Kruskall-Wallis rank sum test (or "one way ANOVA on ranks") is used to quantify the differences between MS proportions at different levels. The p-value smaller than 0.05 for all three proforms ($p < 0.01$) indicates significant differences between the use of proforms at different CEFR levels.

Moving on to the second approach, the multinomial logistic regression yields a 0.77 accuracy overall (95% CI: (0.76, 0.77), $p > .001$). The detailed results in Table 3 show reasonable accuracy statistics for IT microsystems but low recall and precision for THIS and THAT proforms. The difficulty in picking THAT with local context features might come from the higher complexity of the THAT contexts of occurrence due to the form's functional versatility, i.e. subordinator, relativizer, adverbial, demonstrative determiner and proform.

Thirdly, the model based on the neural network yields a 0.74 accuracy (95% CI: (0.73, 0.74), $p > 0.001$). This model shows the same issues as the multinomial logistic regression model regarding the prediction of THIS, although it performs slightly better. However, the performance increases for THAT, with the recall more than doubling. This could be explained by the capacity of the neural network to create a high-dimensional latent feature space, where the different functions of THAT crystallise over different dimensions, disambiguating the use of THAT as proform as a result.

The best performances overall are therefore reached by the neural network approach, although the multinomial regression method offers better performance for IT alone, and the proportions are proved to show statistically significant differences between CEFR levels. In order to leverage the advantages of these three approaches, a possible avenue for future work is to explore a combination of

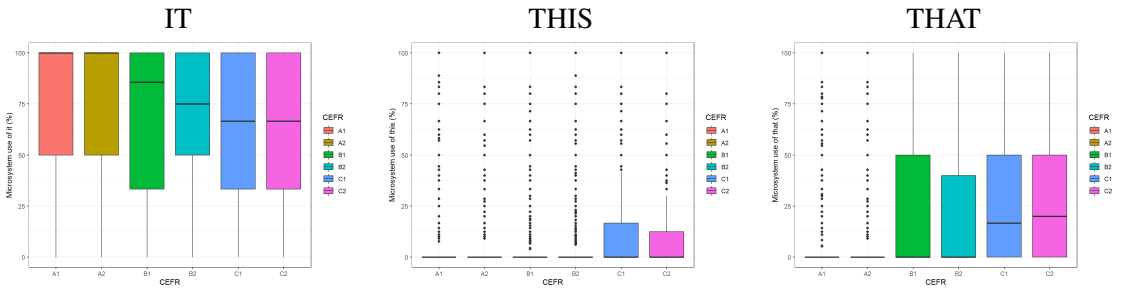

Figure 2: Distribution of relative proportions of IT, THIS and THAT proforms across CEFR levels in the EFCAM-DAT corpus

these models. It could either be a simple concatenation of all measures, leading to a 9 microsystem measures, or a weighted sum of the measures, with weights proportional to the performance of each model on each specific form. Another approach could also lie with bagging several multinomial regressions, and using the average of the output probabilities as our measures.

5.2 Association between measures and CEFR level

We now report the results regarding the models computing associations between the measures and the odds of increasing CEFR levels, using an ordinal regression analysis. Table 4 shows odds ratios for each of the three types of measure. For all measures, the odds ratios are significant ($p < 0.001$), comforting the fact that our microsystem measures can be used as predictors of CEFR level.

These results suggest that writings with higher predicted probability of IT have a reduced odds of being in a higher CEFR level. On the contrary, those with a higher predicted probability of THAT are more likely to have a higher CEFR level. Both predictive methods (multinomial regression and neural networks) agree on a higher probability of THIS being more likely to have a higher CEFR level, while the proportions method finds that on the contrary, a higher proportion of THIS hints at a lower CEFR level. We believe that in this case, this is due to a limitation of the proportion-based measure, that simply counts the percentage of occurrence of each form regardless of linguistic context. The proportion of THIS contains many more outliers than the other two forms, as seen in Figure 2. Our explanation is that the disagreement with the other two models is caused by an inaccurate measure of THIS due to this scattered distribution.

5.3 Discussion

Our first research question revolved around the likelihood of microsystem forms in L2 writings. The three measurement methods we propose capture the choices of a given form with regard to the other possible forms. Regarding model performance, IT is always well predicted while the detection of THIS and THAT could be improved, leading to more accurate probabilities and in turn better microsystem descriptors. To this end, more significant features defining the local context of occurrence of the proforms could be assessed. For instance, adding referential information regarding the degree of givenness of a proform could possibly improve the models. Another way to improve the contextualization of the proforms could be the use of state-of-the-art Natural Language Processing approaches, such as Long-Short Term Memory networks (Hochreiter and Schmidhuber, 1997), or more recently, BERT models (Devlin et al., 2019). Both these methods could be used in the same fashion as we did in the present work, that is, trained to predict a masked form, with the additional benefits of feeding the entire text "as is" to the model and not needing to hand-craft context features. It must be noted that our use of a neural model differs from black-box models that rely on the direct ingestion of texts to predict errors for instance. Our neural model relies on proforms rather than full texts, hence giving specific grammatico-functional probabilities that can be used in subsequent higher-level prediction tasks.

The second research question was to analyse the degree of association between the measures and the CEFR levels given to the texts. Our results indicate that an expert proform MS model can be trained on the basis of likelihood of occurrence, with a slight disagreement between proportion-based and probability-based measures. In both cases, an expert model could use these two mea-

	Multinomial log regression			Neural network		
	IT	THIS	THAT	IT	THIS	THAT
Balanced accuracy	0.71	0.69	0.64	0.74	0.70	0.72
Precision	0.93	0.20	0.36	0.93	0.31	0.66
Recall	0.81	0.54	0.31	0.74	0.70	0.72

Table 3: Performance statistics for the predictive approaches to measuring the proform microsystem

		Odds ratio	95% CI	p_value
Proportions	IT	0.995	0.995, 0.995	<0.001
	THIS	0.997	0.997, 0.997	<0.001
	THAT	1.010	1.010, 1.010	<0.001
Multinom log regression	IT	0.992	0.991, 0.993	<0.001
	THIS	1.006	1.005, 1.007	<0.001
	THAT	1.012	1.011, 1.014	<0.001
Neural network	IT	0. 47	0.42, 0.51	<0.001
	THIS	1.17	1.04, 1.32	<0.001
	THAT	2.27	2.04, 2.53	<0.001

Table 4: Ordinal logistic regression of CEFR by proportion of IT,THIS and THAT

sures as predictors of CEFR levels in new incoming learner writings.

The MS model also supports qualitative feedback with regards to specificity and cognitive fidelity. Firstly, the probability-based models offer knowledge of proform use at word level, allowing specific identification in context, hence specific feedback. A high level of feedback specificity improves understanding from the learner (Shute, 2008). Secondly, because of the grammatico-functional nature of the MS concept, the MS model's measures can be used to explains reasons of a problem. For instance, a proficiency-predicting model relying on MS proform features could point out the demonstrative pronouns in a learner's text in a similar fashion to what Dascalu et al. (Dascalu et al., 2013) do by identifying cohesion gaps. This level of explainability gives a high degree of cognitive fidelity. In this respect, the neural-model increases interpretability as it provides a broader variation of odds ratios, indicating clearer proficiency gaps and making the effects of each form clearer to disambiguate.

6 Conclusion

In this paper, we have reported on the design of new grammatico-functional metrics which are to be used in the expert module of an ICALL system. The metrics rely on paradigmatic syntactic relations between words of specific functions. We have illustrated the approach with the design of the IT, THIS, THAT proform microsystem. The measures rely on the relative proportions of the forms and their likelihood of occurrence. They show significance in terms of association with CEFR levels, paving the way for their introduction in a specific proform microsystem expert model.

7 Acknowledgments

This project is funded by the French National Research Agency. ANR-22-CE38-0015-01

References

Vincent Aleven. 2010. Rule-Based Cognitive Modeling for Intelligent Tutoring Systems. In Roger Nkambou, Jacqueline Bourdeau, and Riichiro Mizoguchi, editors, *Advances in Intelligent Tutoring Systems*, number 308 in Studies in Computational Intelligence. Springer Berlin Heidelberg.

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

John R. Anderson. 2013. The Expert Module. In *Foundations of Intelligent Tutoring Systems*, pages 21–54. Psychology Press.

Yigal Attali and Jill Burstein. 2006. Automated Essay Scoring With e-rater® V.2. *The Journal of Technology, Learning and Assessment*, 4(3):3–29.

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

Christopher Bryant and Ted Briscoe. 2018. Language Model Based Grammatical Error Correction without Annotated Training Data. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 247–253, New Orleans, Louisiana. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

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

Mihai Dascalu, Philippe Dessus, Stefan Trausan-Matu, Maryse Bianco, Aurélie Nardy, Mihai Dascălu, and Ștefan Trăușan-Matu. 2013. ReaderBench, an Environment for Analyzing Text Complexity and Reading Strategies. In *AIED 13 - 16th International Conference on Artificial Intelligence in Education*, volume 7926 of *Lecture Notes in Computer Science (LNCS)*, pages 379–388, Memphis, TN, United States. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Thomas Gaillat. 2022. Investigating the scopes of textual metrics for learner level discrimination and learner analytics. In Agnieszka Leńko-Szymańska and Sandra Götz-Lehmann, editors, *Complexity, Accuracy and Fluency in Learner Corpus Research*, number 104 in Studies in Corpus Linguistics, pages 21–50. John Benjamins.

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).

Jeroen Geertzen, Theodora Alexopoulou, Anna Korhonen, et al. 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.

Yves Gentilhomme. 1980. Microsystèmes et acquisition des langues. *Encrages*, pages 79–84.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Rodney Huddleston and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of The English Language*. Cambridge University Press, Beccles, Suffolk.

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

Anke Lüdeling and Hagen Hirschmann. 2015. Error Annotation Systems. In Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier, editors, *The Cambridge Handbook of Learner Corpus Research*, pages 135–158. Cambridge University Press, Cambridge.

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.

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

Roger Nkambou. 2010. Modeling the Domain: An Introduction to the Expert Module. In Roger Nkambou, Jacqueline Bourdeau, and Riichiro Mizoguchi, editors, *Advances in Intelligent Tutoring Systems*, number 308 in Studies in Computational Intelligence, pages 15–32. Springer Berlin Heidelberg.

Roger Nkambou, Jacqueline Bourdeau, and Riichiro Mizoguchi, editors. 2010. *Advances in Intelligent Tutoring Systems*. Number 308 in Studies in Computational Intelligence. Springer Berlin Heidelberg.

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 and Elena Volodina. 2018. Investigating the importance of linguistic complexity features across different datasets related to language learning. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 49–58, Santa Fe, New-Mexico. Association for Computational Linguistics.

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

Taraka Rama and Sowmya Vajjala. 2021. Are pretrained text representations useful for multilingual and multi-dimensional language proficiency modeling? ArXiv:2102.12971 [cs].

Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. 2018. Second Language Acquisition Modeling. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–65, New Orleans, Louisiana. Association for Computational Linguistics.

Valerie J. Shute. 2008. Focus on formative feedback. *Review of Educational Research*, 78(1):153–189.

Edgar A. Smith and R. J. Senter. 1967. Automated readability index. Technical Report AMRL-TR-66-220, Aerospace Medical Division, Wright-Paterson AFB, Ohio.

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

Anaïs Tack, Thomas François, Anne-Laure Ligozat, and Cédrick Fairon. 2016. Evaluating Lexical Simplification and Vocabulary Knowledge for Learners of French: Possibilities of Using the FLELex Resource. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 230–236, Portorož, Slovenia. European Language Resources Association (ELRA).

Helen Yannakoudakis, Øistein Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for ESL learners.

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