Research Methods in Applied Linguistics 4 (2025) 100238

Contents lists available at ScienceDirect

Research Methods in Applied Linguistics

journal homepage: www.elsevier.com/locate/rmal

Assessing the validity of syntactic alternations as criterial features of proficiency in L2 writings in English

Cyriel Mallart [a,1], Andrew Simpkin [b], Nicolas Ballier [c], Paula Lissón [d], Rémi Venant [e], Bernardo Stearns [f], Jen-Yu Li [a], Thomas Gaillat [a,*]

[a] Université Rennes 2, LIDILE, France
[b] University of Galway, School of Mathematical and Statistical Sciences, Ireland
[c] Université Paris Cité, ALTAE, France
[d] University of Las Palmas de Gran Canaria, IATEXT, Spain
[e] Université du Mans, Laboratoire d'Informatique de l'Université du Mans, France
[f] University of Galway, DSI, Ireland

ARTICLE INFO

Keywords:
L2 english
Proficiency assessment
Syntactic alternation
Complexity measures
Functional analysis
Paradigm competitors

ABSTRACT

This article addresses Second Language (L2) writing development through an investigation of alternation-based metrics. We explore the paradigmatic production in learner English by linking language functions to specific grammatical paradigms. Using the EFCAMDAT as a gold standard and a corpus of French learners as an external test set, we employ a supervised learning framework to operationalize and evaluate seven alternations. We show that learner levels are associated with these seven alternations. Using ordinal regression Modeling for evaluation, the results show that all syntactic alternations are significant but yield a low impact if taken individually. However, their influence is shown to be impactful if taken as a group. These alternations and their measurement method suggest that it is possible to use them as part of broader-purpose Computer-Assisted Language Learning (CALL) systems focused on proficiency assessment.

1. Introduction

Second Language (L2) writing assessment is an essential part of language education. It is a complex task which relies either on human judgment or computer-based measures. Since Page's first system (1968), assessing writing proficiency with computers has generated much interest. Historically, such systems have always relied on automatic measurements used as sets of features to represent the writing construct. Based on these representations regression models have been employed to predict proficiency. Due to the central role of features, recent debate has focused on the tension that exists between their predictive power and their explanatory value. On the one hand, many features belong to the Complexity, Accuracy and Fluency (CAF) framework and have been shown to present high predictive power in modeling proficiency (Kyle, 2016; Kyle et al., 2021; Vajjala, 2018). On the other hand, Biber et al. show that many CAF measures are "omnibus" in that they "collapse consideration of multiple structural and syntactic features into a single variable"(Biber et al., 2020). They advocate for measures that target specific structural forms mapped to specific functions as part of descriptive models (Biber & Gray, 2011; Staples et al., 2016). Nonetheless, both approaches rely on the syntagmatic order of language (Halliday & Matthiessen, 2004, p. 24) in that tokens or lexico-grammatical patterns are identified and combined

* Corresponding author.
E-mail address: thomas.gaillat@univ-rennes2.fr (T. Gaillat).
[1] New affiliation: INRIA Rennes, Campus de Beaulieu, 263 Av. Général Leclerc, 35042 Rennes, France.

https://doi.org/10.1016/j.rmal.2025.100238
Received 2 April 2025; Received in revised form 3 July 2025; Accepted 3 July 2025
Available online 4 August 2025

along the syntagmatic axis following a count-based approach. This is an essential dimension, but it could be complemented by measures focusing on the second dimension of language which is the paradigmatic order, i.e., word or construction choices in a particular position. In other terms, automated systems could benefit from extra measures dedicated to potential paradigmatic substitutions between words of the same category. We know that learners select forms to realize specific functions in the L2 (Ellis, 1994, p. 374); in doing so they hesitate between sets of forms at particular positions. It could be helpful to have measures that capture such paradigmatic substitutions and help understanding the functions they represent. Modeling L2 proficiency could benefit from a combination of syntagmatic and paradigmatic features.

Our study is grounded in the framework of alternation research studies. The aim of our approach is to complement CAF-related features in L2 modeling. We focus on how learners' "alternate ways of 'saying the same thing'" (Labov, 1972). We are particularly interested in syntactic alternation in L2 English as it gives an insight into the paradigmatic choices that a learner can make in a particular position. Example 1 shows a learner error which may be the result of hesitations in using article *the* instead of possible competitors such as *a* or article 0.

(1) * "What do you think about positive discrimination in *the* companies?" (EFCAMDAT writing ID: 569744)

Previous work has focused on linking contextual features to variants in an attempt to explain the influential factors in learners' choices of variants (Gries & Wulff, 2013). Results have shown that learner choices depend on various types of factors involving local and global features ranging from lexico-grammatical to semantic dimensions. However, these studies did not address the link between alternations and proficiency. One notable exception includes the analysis of the genitive alternation in L2 English learners (Dubois et al., 2023), in which proficiency is used as a predictor variable when modeling for the *of*- or *s*-genitive variants. In our case, we are interested in how variant usage is linked to proficiency. Our purpose is to investigate how alternations can be used to model proficiency as an outcome. The long-term objective is to design alternation-based metrics for proficiency prediction.

We argue that it is important to understand what forms learners hesitate with when writing their sentence, because these competing forms may account for some specific L2 uses. Measuring the evolution of probabilities of usage for a certain form across proficiency levels could help us analyze how learners improve over time, investigating how these forms compete with each other. This could cast new light on how learners tend to favor one form over the others, depending on context. Our assumption is that the forms' variations can be associated with developmental patterns in writing. Indeed, writings may exhibit variations corresponding to specific developmental stages and proficiency levels. For some aspects of the linguistic system, developmental stages do not imply an increase in linguistic complexity. For example, compounds are cognitively more complex and appear later in proficiency levels. A paradigmatic approach is not simply linked to complexity but it also analyzes which form is used rather than another one. In a previous study, alternations, conceptualized as microsystems, were analyzed in terms of relative proportions of forms belonging to the same linguistic paradigms (Gaillat et al., 2022). It was shown that such proportions were associated with proficiency but lacked adaptability to contexts.

Our proposal is to quantify alternations using the probability of one variant relative to its counterparts. To evaluate the discriminating potential of alternations in the measurement of proficiency, we assess how the probability distributions of variants can be associated with proficiency. The rest of the paper is structured as follows: Section 2 lays out the theoretical background. Section 3 is dedicated to presenting our method. In Section 4 we present and analyze the results. Section 5 focuses on the discussion and Section 6 concludes on the findings and future perspectives.

2. Theoretical background

2.1. Assessing writing ability and L2 proficiency

The construct of writing ability is broad as it covers several underlying purposes. It is essential to distinguish between learning to write and writing to learn in areas other than language learning. Cushing Weigle (2013) refines this distinction by identifying three purposes for assessment. Firstly, she considers Assessing Writing (AW) as a way to verify if students have skills in text production including revisions and pragmatic aspects. Secondly, she points out that Assessing Content through Writing (ACW) verifies whether students understand specific content. Finally, she defines the task at hand in this paper using Assessing Language through Writing (ALW). This task addresses whether students master "the second language skills necessary for achieving their rhetorical goals in English" (Cushing Weigle, 2013) or not.

The purpose of ALW is to assign some level of proficiency to a written production in a foreign language. In order to establish an association between production and proficiency scientifically, some methods rely on analytical rubrics (Knoch, 2011), others on fine-grained checklists (Safari & Ahmadi, 2023). In the case of Automatic Essay Scoring (AES), recent approaches of ALW have relied on Large-Language Models (LLM) including non-explicit features (Banno et al., 2024; Mizumoto & Eguchi, 2023; Yamashita, 2024; Yancey et al., 2023). Conversely, traditional approaches to AES have relied on explicitly selected features that have been validated as being criterial for proficiency (Hawkins & Filipović, 2012). In this case, many features have been sought within the linguistic complexity framework (Bulté & Housen, 2012) including measures of lexical and grammatical diversity (for a recent overview see Kuiken (2023)). As a result, linguistic complexity features, and their measures have played a central role in automatic proficiency assessment systems without which the writing ability construct may neither be measured, nor linguistically substantiated.

Recent developments have focused on the linguistic choices that learners make in particular contexts in terms of syntagmatic associations. Granger and Bestgen (2014) used the strength of association between adjective and noun combinations to model learner

proficiency. Similar studies followed to model proficiency using the strength of association between words in particular grammatical relationships such as verbs and objects (Eguchi & Kyle, 2023; Paquot, 2019). Researchers have also modeled proficiency based on verbs and their syntactic environments (DeVore & Kyle, 2023).

Most if the approaches rely on syntagmatic features, meaning that many complexity metrics rely on sequential feature counts in texts. This essential dimension could be complemented with features reflecting the paradigmatic choice that presents itself to learners when unfolding their discourse. This is where alternation studies may contribute to proficiency assessment.

2.2. Syntactic alternation in L2 studies

Alternation can be defined as "structurally and/or lexically different ways to say functionally very similar things" (Gries, 2017). In the case of syntactic alternation, the framework has served functional approaches by linking variants to functions in terms of paradigmatic relations. These relations provide an insight into the production choices available in discourse. A paradigmatic analysis helps identify which options a speaker has at a given moment in their unfolding discourse. Paradigms are particularly interesting as they usually group variants according to specific language functions. For instance Bresnan studied the dative structure in native English and the variation between ditransitive and prepositional structures (Bresnan et al., 2007). Not only did their work formalize the dative as a paradigmatic structure, but it also gave empirical evidence of the contextual features leading to the choice of one or the other variant. Their approach was not long before being applied to L2 language, especially in the context of approaches intersecting corpus linguistics with statistical Modeling.

To date, many different linguistic systems have been analyzed within this framework, starting off with the dative alternation (Jäschke & Plag, 2016). Research efforts then focused on other forms including particle placement (Kinne, 2020; Paquot et al., 2019), *that* complementation (Wulff et al., 2014), modals (Gries & Deshors, 2014), proforms (Gaillat et al., 2022), prenominal adjective order (Wulff & Gries, 2015), preposition–article contractions in Portuguese L2 (Picoral & Carvalho, 2020) and the genitive (Gries & Wulff, 2013). These studies analyzed variants in their contexts of occurrence, trying to identify various morpho-syntactic and semantic factors as well as psycholinguistic factors such as priming, similarity and surprisal. In these approaches, the variants were extracted whether they were used correctly/appropriately or not.

Usage-based approaches to Second Language Acquisition hinge on the form-meaning pairing (Ellis, 1994) in which learners learn to associate constructions to their meanings. The learning process partly relies on implicit exemplar-based learning mechanisms in which the learner is confronted with several alternatives to associate a function to a form (Wulff & Ellis, 2018). In doing so, the learner learns to detect the constraints that trigger a construction. Depending on the acquisition stage and following the Competition Model (MacWhinney et al., 1984), the learner may have several forms in mind that compete for the realization of a meaning. This competition reflects the variable nature of Interlanguage. Low-proficiency learners, more prone to L1 transfers, may be less sensitive to linguistic constraints, leading to more inappropriate or inaccurate forms in groupings than with higher-proficiency learners. As a result, alternations in L2 analysis may include inaccurate or inappropriate groupings of forms per function. The construct captures the likelihood of choosing a form depending on linguistic constraints whether they are accurate or not. In this respect, the alternation construct is an observation of the likely alternatives that a learner has when selecting a construction to express a function. It may be that these alternatives vary with proficiency. The aforementioned studies showed fruitful results in the description of L2 subsystems linked to specific linguistic functions. However, to the best of our knowledge, the variants were not looked at in relation to proficiency.

This is what was achieved by Dubois et al. (2023) who analyzed the genitive in an L2-English learner corpus annotated with CEFR levels. Their findings showed that lower-proficiency learners differed from native and higher-proficiency learners. Learners appeared to be less sensitive to fine-grained features such as definiteness and animacy on the proficiency continuum. In sum, the authors used proficiency as a predictor of the variants. Gaillat et al. (2022) also included proficiency in their experimental design. However, instead of modeling alternations – which they called microsystems – they modeled levels of the Common European Framework of Reference (CEFR) (Council of Europe, 2018) as a function of proportions of alternation variants. Their results showed evidence of proportion-dependent variations across CEFR levels. These findings suggests that alternations may be criterial features of L2 development (Hawkins & Filipović, 2012). However, this approach relied on actual word choices and ignored the contexts of occurrence triggering form use. So how could alternations be operationalized with contextual information to address the question of their association with proficiency?

Hesitations in word choices can be captured at text level by observing how likely learners are to use variants in expressing a function. A way of approaching the task is to consider the probability of occurrence of one form versus the others that map to the same function. In this respect statistical methods such as Generalized Linear Models (GLM), including logistic regression, provide solutions to calculate probabilities of using each form based on the local context. Gries and Deshors (2014) set up the MuPDAR (Multifactorial Prediction and Deviation Analysis using Regressions) approach, a multifactorial regression analysis. As an illustration, they analyzed the probabilities of *may* or *can* in a learner corpus based on a model trained on native English. The predicted probabilities of the modals were used to determine whether the learners made a canonical choice or not. Although they acknowledged proficiency as potentially interfering with learners' linguistic choices, the authors did not operationalize proficiency as a variable.

Specific syntactic alternations could be exploited in the assessment of L2 writing with regression-based methods using proficiency as the outcome variable. However, the empirical validity of such an approach calls for a number of requirements before using it for the description of writing development. First, specific alternations need to be functionally identified. Secondly, a measurement method needs to be operationalized. Finally, it must be evaluated in terms of how the measurements relate to actual L2 proficiency. One way of representing proficiency is to use the CEFR. This approach offers the advantage of linking learner proficiency to a scale that is widely used by practitioners. Using syntactic alternations for proficiency assessment raises three research questions:

Table 1
Writings across CEFR levels in the EFCAMDAT and CELVA.Sp corpora.

Writings	# of writings		% of writings		average # of words		Standard Deviation of words	
CEFR	EFCAMDAT	CELVA.Sp	EFCAMDAT	CELVA.Sp	EFCAMDAT	CELVA.Sp	EFCAMDAT	CELVA.Sp
A1	341,155	90	47.16	8.61	39.07	141.43	14.41	83.16
A2	215,344	324	29.77	31.00	64.53	206.40	17.91	99.03
B1	116,539	358	16.11	34.26	94.75	261.16	21.47	124.59
B2	40,238	212	5.56	20.29	134.86	319.93	33.22	141.44
C1	10,006	53	1.38	5.07	169.34	398.70	26.59	161.94
C2	NA	8	NA	0.77	NA	388.63	NA	152.04
Total	723,282	977	100	100	62.75	253.85	34.87	135.40

1. Which L2 syntactic alternations could be mapped to linguistic functions?
2. How can these linguistic functions be modeled in terms of alternation probability ?
3. What is the relationship between observed alternation probabilities and learner language proficiency?

3. Methods

In this section, we describe the corpora, the processing pipeline including the operationalized alternations.

3.1. Corpus data

Because the purpose of the approach is to mimic syntactic alternations in learners, we choose two learner corpora. One is used for training and internal testing purposes; the other one is used for external evaluation. The first corpus is the EFCAMDAT (Geertzen et al., 2013) in its refined version as described by Shatz (2020). The refined version of the EFCAMDAT is a collection of 723,282 writings collected online by *Englishtown* language schools across eleven countries. Metadata about the learners include their L1 and their proficiency level. The learners were required to write texts following prompts such as "introducing yourself by email" and "writing a movie review". Due to the variety of writing prompts, the types of genres were not controlled. Regarding the assignment of CEFR levels, the authors of the corpus established a matching table (Geertzen et al., 2013, p. 241) between the 15 learners' skill levels of *Englishtown* and external proficiency scales including the CEFR . The refined version of the corpus does not include C2 writings. *Englishtown* relies on language teachers to manually correct and grade the writings, which allows them to gradually move from one skill level to the next. It should be noted that inter-rater agreement was not reported regarding the consistency of grades between texts. The proficiency levels attributed to the texts actually correspond to the successful completion of coursework levels of *Englishtown* by these students. The completion of each *Englishtown* level is used as a proxy of their acquired skills. Table 1 provides descriptive statistics of the distributions. In our study, the corpus is split between training and test sets for several Modeling tasks applied to different alternations.

The second corpus is used as an external validation dataset in order to evaluate the generalization potential of our analyses. The *Corpus d'Étude des Langues Appliquées à une Spécialité* (CELVA.Sp) (Mallart et al., 2023) was collected in two French universities and includes 977 writings produced by undergraduate and post-graduate students (see Table 1). These students, aged 19 to 24, are enrolled in courses in several domains ranging from mathematics to sport and pharmacy. The corpus metadata[2] includes different types of behavioral information such as their exposure to L2. The writings were annotated by four language certification experts[3] who followed a protocol based on the descriptors of the writing production competence of the Common European Framework of Reference (CEFR) (Council of Europe, 2018, Appendix 4, p. 187–189).

To evaluate inter-rater reliability, we randomly and sequentially extracted two samples of 30 texts each that were annotated independently by the four annotators. An annotation-adjustment discussion session was conducted between the two samples. The Kappa results obtained for the first sample showed values ranging from .52 to .79. Table 2 shows the confusion matrix between the raters. The second sample also showed fair to good agreement as per Fleiss (2003), yet less than the 0.8 value mentioned by Artstein & Poesio (2008). The lower bracket of the agreement values in the confusion matrix points to one annotator who appears to consistently disagree with the three others. To test the significance of this disagreement, permutation tests between the first and second sample were conducted. Results showed no significant difference (p-values >.05 for all pairwise kappa differences). These results must be read in the light of the complexity of the task which involved classifying entire texts into one of five categories among four annotators. This clearly leads to a more difficult task for interpretation. Given the experience of the experts and the CEFR rubric that was used, agreement levels are acceptable. Each individual annotator was then given a split of the remainder of the corpus to annotate.

[2] The corpus and metadata files are available online from https://nakala.fr/collection/10.34847/nkl.5f8an0ke.

[3] These raters are teachers of English as a Foreign language with more than 20 years experience in the language centers of the two French universities. They have extensive experience in the national language certification examination (CLES) as well as other certifications such as the Cambridge and the TOEIC tests in higher education.

Table 2
Inter-rater agreement for a CEFR annotation task conducted by four raters on 30 writings sampled from the CELVA.Sp corpus.

Raters' pairwise agreement (Cohen's Kappa)	1	2	3	4
1	–	–	–	–
2	.52	–	–	–
3	.79	.61	–	–
4	.76	.55	.75	–

Table 3
The seven alternations considered for this study.

alternations	variants	Function	Examples of confusions
Proforms	it, this, that	reference to entity	The student cares for this/that/it
Multi-noun	compound, genitive, prepositional	Pairs of nouns functioning as compounds, genitive or prepositional phrase	She took a student loan/a student's loan/the loan of a student
Articles	a, the or 0	determining a noun	a/the/0 loan
Duration	for, since or during	complementing a verb with duration related information	The student has had this loan for/since/during 2 years.
Quantifier 1	any, some	determining a quantity: one or more or unspecified respectively	Any/some students could help.
Quantifier 2	many, much	determining an important quantity	Many/much hard-working students do not rest.
Relativizer	that, which, who	surbordinator refering to entity	The students who/that/which study.

3.2. Defining potential alternations and operationalising their measurement

In a previous study, several alternations were identified and tested in terms of proficiency level (Gaillat et al., 2022). In our study, we use seven of these alternations for which we suspect potential acquisitional confusion. They refer to specific syntactic or semantic functions and are described in Table 3. For instance, when referring to an entity or a whole clause, learners do understand the need to use a proform but they may get confused between IT, THIS or THAT, leading to a potential semantic error. Learners may also get confused between relativizers, leading to a syntactico-semantic error.

To operationalize the construct, we adopt a different approach from a previous study based on counts and proportions (Gaillat et al., 2022). As an alternation represents text at local level in terms of constructions, it is possible to model its variants in terms of probability of occurrence. In this paper, we propose a supervised learning method based on multinomial logistic regression as in (1). Each alternation Y_i for texts $i = 1, \dots, n$ can take any one of K forms (e.g. proform can be IT, THAT or THIS so $K = 3$), and is assumed to follow multinomial distribution with parameter $\pi_i = (\pi_{i1}, \dots, \pi_{iK})$. In this framework, the probability of Y taking some discrete value k ($k = 1, \dots, K$) is modeled as a function of P predictor variables $X_1, \dots, X_P$. The following equation gives the probability that a response is equal to some class K conditional on the covariates.

$$Pr(Y = k|X) = \frac{exp(\beta_{0k} + \beta_{1k}X_1 + \cdots + \beta_{nk}X_n)}{1 + exp(\beta_{0k} + \beta_{1k}X_1 + \cdots + \beta_{Pk}X_P)} \quad (1)$$

The model returns values indicating the predicted probability of using each form of an alternation. The explanatory variables $X_1, \dots, X_P$ are the features extracted from the local syntactic context in which the forms appear.

3.3. Feature extraction

To extract the features, we process the data in two stages. First, we conducted automatic annotation of the entire corpus with the use of UDPipe (Straka et al., 2016). UDPipe provides multilevel annotation relying on the Universal Dependency (UD) framework v2.0 (de Marneffe et al., 2021). Each token is associated to linguistic information ranging from Parts of Speech (POS) to UD and includes morphological features such as person and number. These morphological features were specific to UDpipe and motivated the authors when designing the experiment in 2023. The authors recognize that other parsers could have been used.

Secondly, where the QUANT pattern is the lemma MANY used as an adjective modifying the noun it precedes. This includes cases where there are adjectival and adverbial modifiers in between. Formulaic use is not distinguished from other production choices.

$$QUANT[lemma = \text{“}many\text{”}]; N[upos = NOUN]; N - [amod] - > QUANT; \quad (2)$$

Table 4
Features used for the proform alternation model.

Feature type	Feature description
POS	Left context 3-gram POS (UD and PTB)
POS	Right context 5-gram POS (UD and PTB)
POS	POS of dependency head
Dependency	Head-dependency relation between form and head
Dependency	Normalized dependency distance to root
Tokens	Number of tokens in pattern
Morphology	Next token number
Morphology	Next token person
Morphology	Next token mood (UD)
Morphology	Next token verb tense
Nationality	Nationality declared by learner

We defined patterns for all the variants of the seven alternations. Applying them to texts yielded an alternation-specific dataset made up of the targeted variants together with their linguistic features given by UDpipe. Each data set was a representation of a function mapped to its variants. For instance, each POS,[4] UD relation and morphological feature appears in dedicated feature columns. Note that not all morphological features apply to all types of tokens, resulting in cases where some morphological features are null. Likewise, some target tokens may be located near sentence boundaries leading to right or left context features being absent, thus null.

3.4. Feature selection

Prior to Modeling, we conducted feature selection. First we removed features with more than 50% missing values among all texts. A lower threshold would have lead to the removal of features related to the target being at the beginning and ending of sentences (due to non-existing left or right context in each sentence). This was to avoid removing features where an alternation occurs as the first or last word of a sentence. Conversely, a higher threshold would include more features but remove more observations in the Modeling phase, and vice-versa. We found this threshold struck a good balance. We also dropped UDpipe features that actually describe the forms such as *lemma*, *wordform* and *textform* as these would trivially explain the occurrence. Table 4 lists all the selected features for the proform alternation (see Appendix A for the feature set of each alternation).

3.5. Classifying the forms

To perform classification we employed multinomial logistic regression on the EFCAMDAT training data and predicted labels with a testing subset. We first randomly split the data into 80% training and 20% testing. The random sampling occurred within each class and preserved the overall distribution of the data. Secondly, as the training set was imbalanced in terms of forms (less forms of one type than its competitors), we randomly subsampled the set to the lowest number of the variants making up a given alternation. For instance, there are fewer occurrences of THIS (34,484) compared to IT and THAT in the corpus. Hence, we selected 34,484 instances of each of the IT and THAT proforms at random before training the model. Subsampling was necessary to train a model without the bias of frequency. Variant IT being most frequent would lead the model to favor it, hence tempering the impact of contextual features. Conversely, the test set was preserved as its imbalanced nature reflected the natural distribution of the variants.

To predict the forms in context, we fitted the multinomial regression model on the training data made up of syntactic alternation variants (outcome variable) and their contextual features. As the model omits observations with missing values, there were 20,063 observations in the PRF model.[5] We then applied it to the EFCAMDAT testing data (see Section 3.6 to obtain probabilities of occurrence of each variant in each slot. In other terms, the model predicts a form on the basis of the contextual features. These predictions reflect what a learner would choose.

3.6. Evaluation method

Evaluation required a two-stage process involving i)the creation and evaluation of alternation predictive models, and (ii) using the EFCAMDAT to compute alternation probability scores and evaluate their associations with proficiency levels. Fig. 1 illustrates the process.

Annotation First, we prepared a GS to evaluate how our tools would map variants to specific functions. We used the external corpus (see Section 3.1) to create a subset for each alternation. We randomly extracted seven subsamples of circa 250 occurrences, and three linguists[6] annotated whether the forms matched the definitions listed in the annotation guidelines previously prepared

[4] This includes UD POS and Penn Tree Bank (PTB) POS.

[5] There were 1373 in the DUR model, 37,630 in the MULTINOUN model, 12,844 in the QUANT1 model, 6499 in the QUANT2 model, 9984 in the REL model.

[6] Two doctors in linguistics (the annotators) carried out the annotation, and a professor of linguistics (the consolidator) validated the cases in which differences were identified and corrected according to the annotation guidelines.

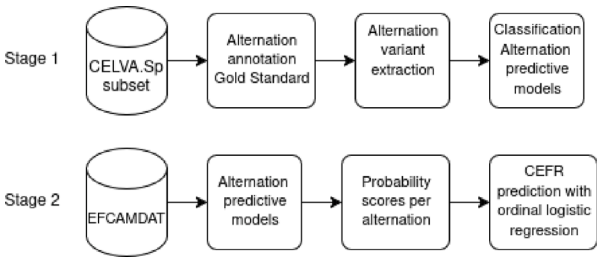


Fig. 1. Two stages to evaluate the alternation models.

(See Appendix B for annotation details including guidelines). In other words, the subsets only included the forms that were mapped to the same function. We computed inter-annotator agreement with the Fleiss' Kappa index (Fleiss et al., 2003). The choice of this index was motivated by its ability to deal with categorical data and to correct for chance agreements (Larsson et al., 2020). Disagreements were treated by the third linguist in charge of consolidation.

Extraction Based on the Gold Standard, we could then evaluate the extraction queries described in Section 3.3. This gave an indication of how robust the queries were and the quality of the mapping between extracted forms and the alternation system they were supposed to be part of. We applied the queries to the same external corpus subsets as the GS and computed accuracy metrics (F1-Score, precision and recall). This yielded information on the quality of the extraction of the syntactic alternation variants to see if we could apply them to extract variants from the entire training and tests sets for the classification task.

Classification The prediction performance of the alternation models was evaluated at the word level on the test set of the EFCAMDAT corpus. We compared predicted variants with the actual forms used by learners. As the test sets were imbalanced, we chose to report balanced accuracy.

Evaluating syntactic alternation associations with proficiency Finally, we performed two tests to analyze prediction distributions. A Kruskal–Wallis rank sum test allowed us to analyze differences between CEFR levels. We also used ordinal logistic regression to investigate whether an association between the predicted probabilities of alternation use were associated with the odds of increasing the CEFR level. In the latter case, form probability distributions were aggregated at text level with the median. Odds ratios indicated potential effects of the increase of form usage on proficiency. The evaluation was conducted on the test sets of the EFCAMDAT and CELVA.Sp corpora.

4. Results

4.1. Alternation annotation

As a preliminary step to evaluate extraction, we built a GS to test how well candidate alternation forms (described in Table 3) could be identified according to the linguistic functions they mapped. Human annotation of seven syntactic alternation-specific subsets revealed a very high level of agreement between the two annotators (Table 5) including Fleiss' Kappa values mostly above 0.9. These results showed that mapping the forms to their linguistic functions casts little doubt among expert annotators. However, some differences remained, and these were treated individually. Apart from obvious issues due to cognitive tiredness, differences were due to ambiguities in learner language. Example 2 illustrates this issue with confusion around the use of THAT. The THAT proform annotation guidelines indicate: "Annotate THAT only as proform, not as determiner, nor adverbial, nor relativizer nor complementizer". However, in this example the context of occurrence was ambiguous as the learner did not insert the obligatory *it or that* leaving an ambiguity in interpreting the used *that* as either a proform in subject position or a complementizer. In these cases the consolidator advocated for not tagging the forms as proforms and for applying this to all similar cases in order to provide consistency. The annotation differences are listed in Appendix C.

(2) a. "My opinion in the invention on the web is *that* is allowed at the time to start to communicate more easily, to exchange document ".

b. "Except *that* is also important to consider the negative outcomes we can get from it".

By treating all the differences individually, decisions were made according to each context by respecting coherence in their application as in the case for proform or complementizer *that*. This process resulted in a consolidated dataset used as GS in the remainder of the experiments.

Table 5
Inter-annotator agreement for each of the alternations.

Alternations	N	Fleiss Kappa	z	*p*-value
A THE ZERO	160	0.938	19.36	0.000
IT THIS THAT	165	0.951	21.06	0.000
MUCH MANY	110	0.937	12.160	0.000
MULTINOUN	135	0.959	18.64	0.000
SINCE FOR DURING	165	0.886	19.545	0.000
SOME ANY	109	0.985	13.495	0.000
WHICH WHO THAT	165	1	22	0.000

Table 6
Quality of alternation extractions in the GS dataset.

Alternations	Support	Precision	Recall	F1-score
A THE ZERO	160	0.77	0.79	0.77
IT THIS THAT	165	0.87	0.87	0.86
MUCH MANY	110	0.79	0.87	0.77
MULTINOUN	135	0.71	0.76	0.72
SINCE FOR DURING	165	0.83	0.83	0.82
SOME ANY	109	0.75	0.77	0.74
WHICH WHO THAT	165	0.88	0.88	0.87

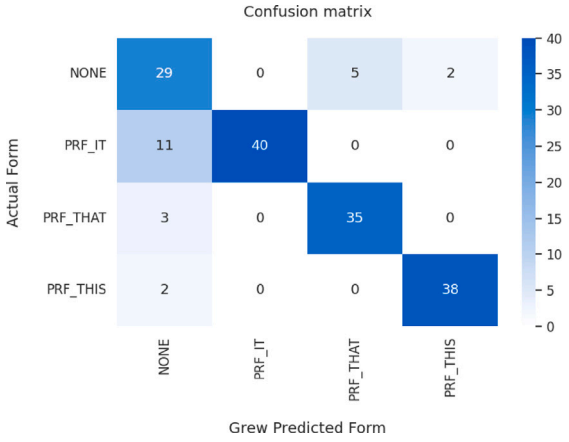


Fig. 2. Confusion matrix for the extraction of IT, THIS and THAT proforms in the Gold Standard dataset.

4.2. Alternation extractions

To evaluate feature extraction and more specifically the quality of the Grew queries, we applied the query tool to the GS made up of annotated sentences and identified form-function mappings (see Section 3.3). Results show how well syntactic alternation forms were extracted by the query tool (see Table 6). The F1-score appears to be above 0.7 for all MS, showing a satisfactory level of robustness for wide-scale extractions. This is confirmed by confusion matrices as in Fig. 2 showing proform extractions.

Precision and recall results show a good balance, indicating no strong bias towards either missing correct forms or badly identifying forms. Nevertheless, some issues remain regarding the A/THE/Ø alternation. The Grew query does not capture the Ø article very well (in front of nouns), hence a rather low F1-score. Appendix D gives details about the extraction results, including itemized accuracy metrics of other alternation variants.

4.3. Classification of alternations

Knowing how well the query tool performed on the GS, we applied queries to the entire training set and the EFCAMDAT test set. After conducting feature selection (see Section 3.4) we modeled the use of syntactic alternations as an outcome variable using local context features as predictors. The purpose was to obtain probability scores to be subsequently used for CEFR classification. Applying multinomial logistic regression for classification, we obtained accuracy measures for each of the alternations in the EFCAMDAT test set. Prediction performance of forms in context appeared to be a challenging task. Table 7 shows the consolidated results. For instance, we can see how well the local context can correctly predict the use of IT, THIS or THAT. For each proform, balanced accuracy shows that more than 70% of cases are predicted correctly by their local context features. Nevertheless, precision shows that THAT and THIS proforms tend to be mistagged. Similar results can be observed for other alternations.

Table 7
Results for the alternation classification with the EFCAMDAT test set.

Alternations	Global accuracy (95% CI)	Balanced accuracy	Recall	Precision
A THE ZERO	.91 (.910, .911)	.875/.852/.997	.822/.740/.996	.669/.864/.999
IT THIS THAT	.67 (.667, .673)	.734/.701/.712	.692/.563/.583	.932/.291/.240
MUCH MANY	.87 (.866, .876)	.869		
MS MULTINOUN {N2 N1/N1ofN2/N1'sN2}	.56 (.559, .563)	.657/.702/.700	.502/.677/.693	.828/.523/.142
SINCE FOR DURING	.73 (.725, .743)	.823/.795/.751	.726/.743 /.681	.571/.941/.309
SOME ANY	.82 (.824, .832)	.817		
WHICH WHO THAT	.6 (.598, .614)	.732 /.670/.710	.656/053/.637	

4.4. Associations with CEFR levels

This section reports the main findings of the study. Using the trained models for each alternation, we obtained predictions for all the occurrences of the alternation variants in the internal and external test sets. Here, we investigated whether these predicted probabilities were associated with proficiency. We focus on the proform alternation as an illustration. Fig. 3 shows the variations of the median probabilities of each proform per text across the CEFR levels in the EFCAMDAT. The probabilities of IT seem to decrease as proficiency increases with significant differences between each level (Kruskal–Wallis rank sum test p-value <.01). THAT seems to be trending in the opposite direction, while THIS shows a slight variation for level A2 (Kruskal–Wallis rank sum test p-value <.01 in both cases). Plotting all alternations shows unequal levels of variations depending on alternations. The quantifier SOME-ANY alternation shows stark variations, while the relative pronoun alternation reveals quite similar medians across CEFR levels of the EFCAMDAT.

Comparing probability distributions between the two test sets shows differences in several alternations. This can again be illustrated with the proform alternation (see Fig. 4). Predictions between CEFR are not as clear-cut in the CELVA.Sp dataset. For instance, the Kruskal–Wallis rank sum test reveals that, while probabilities are significantly different in the EFCAMDAT between CEFR groups (N = 93,072, p <.001 for each proform), it is not always the case in the CELVA.Sp (N = 905, p <.05 for THIS but p >.05 for IT and THAT). A close analysis of other alternations reveals similar contrasting results, which suggests opposite trends in several cases. These differences might stem from the type of corpus data, including writing task and types of learners (see Section 4.5). Nonetheless, consistent results between the two datasets suggest that some variations are corpus independent, indicating that some syntactic alternations help discriminate proficiency.

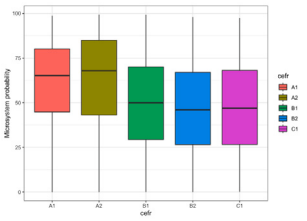
Finally, we performed ordinal logistic regression to investigate whether there was an association between the predicted probabilities of alternation use and the odds of increasing CEFR level. Table 8 shows the results obtained for the EFCAMDAT set. Odds ratios indicate the odds of a higher CEFR level for every 1% increase in the probability of use of one of the variants of an alternation. For instance, for a 1% increased probability of IT, the chances of improved proficiency drop slightly (.995 < 1). Conversely, for every unit increase of THAT probability, the chances of improved proficiency increase (1.011 > 1). In these cases, this would suggest that the probability of proform THAT vs IT and THIS tends to favor better proficiency. Similar observations can be made for MANY, N of N structures, DURING and SINCE, SOME, and finally WHO and WHICH. All of these forms tend to indicate better proficiency in the EFCAMDAT. This means that the more likely a learner context triggers a variant, the more it has a positive or negative impact on proficiency. The alternation model acts as a proxy of the learner's competence in a given alternation microsystem.

The same reasoning can be applied with the odds ratios obtained with the external test set. Some of the observations made on the EFCAMDAT remain the same, and they are shown in bold in Table 8. However, some other findings show opposite trends in the CELVA.Sp. We discuss this in Section 5.

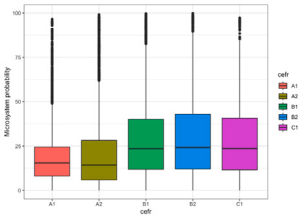
We also conducted feature importance analysis with the varImp method in R's caret (Kuhn, 2015) package (see Table 9 for details on the types of POS present in the right and left contexts of proforms). We computed the scaled importance (Mizumoto, 2023), i.e., what percentage of the model each feature is responsible for. Important features vary according to each alternation and the percentage distribution of its variants spreads considerably. For the proform alternation, the largest feature percentage was 2.93% for the possessive in the right context of the form with a 5 word window. The results concerning the seven alternations are available online.[7]

Overall, we can make some common comments regarding feature importance across all alternations. Most important features were Penn-treebank POS tags as opposed to Universal POS tags also used in the models. This suggests that finer-grained morpho-syntactic annotation helps the classifiers. Finally, analysis of the top-10 features suggest that alternations mostly rely on previous-context features (i.e., multinouns, quantifiers ANY/SOME,) while others mostly rely on post-context features (i.e., proforms). For others (i.e., quantifiers MANY/MUCH, relativizers and duration) both contexts were important.

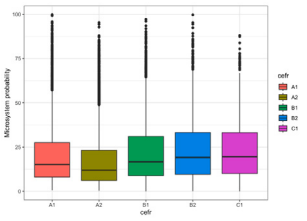
[7] https://www.iris-database.org/details/FmaEH-iw8GR



(a) Median probabilities of IT.



(b) Median probabilities of THAT.



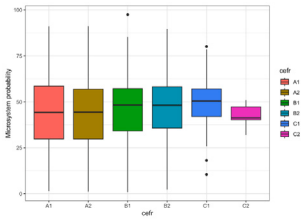
(c) Median probabilities of THIS.

Fig. 3. Distribution of median probabilities of proforms per text in the EFCAMDAT test set.

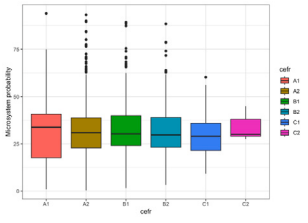
Table 8
Odds ratio between alternations and CEFR levels in the EFCAMDAT and the CELVA.Sp.

Alternations	variants	EFCAMDAT		CELVA.Sp	
		Odds ratio	95% CI	Odds ratio	95% CI
DET	A	.998	.997, .998	1.002	.988, .1.015
	THE	1.004	1.003, 1.004	1	.987, .1.014
	0	**.996**	.996, .996	**.999**	.996, .1.003
PRF	IT	.995*	.995, .996	1.006*	1, 1.013
	THAT	1.011*	1.011, 1.012	.998	.99, 1.006
	THIS	**.998***	.997, .998	**.974***	.962, .986
MLTNN	N2 N1	**.993***	.993, .993	**.989***	.976, 1.001
	N2 S N1	**.999***	.999, 1	**.995**	.984, 1.005
	N1 OF N2	**1.006***	1.006, 1.007	**1.004**	.996, 1.013
DUR	DURING	**1.005***	1.004, 1.007	**1.004**	.995, 1.012
	SINCE	1.008*	1.007, 1.01	1	.992, 1.007
	FOR	**.99***	.989, .991	**.991***	.978, 1.004
QUANT1	ANY	.983*	.982, .983	1.004	.998, 1.011
	SOME	1.018*	1.017, 1.018	.996	.989, 1.002
QUANT2	MANY	1.012*	1.011, 1.013	.995	.989, 1.001
	MUCH	.988*	.987, .989	1.005	.999, 1.011
REL	THAT	**.992***	.99, .993	**.979***	.963, .995
	WHO	1.001	.999, 1.002	.991*	.983, .999
	WHICH	**1.009***	1.007, 1.01	**1.011***	1.003, 1.02

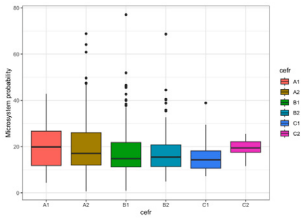
* *p*-value <.05.



(a) Median probabilities of IT.



(b) Median probabilities of THAT.



(c) Median probabilities of THIS.

Fig. 4. Distribution of median probabilities of proforms per text in the CELVA.Sp external test set.

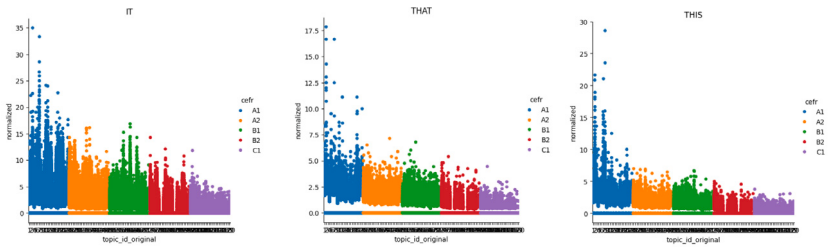
Table 9
Top features used for the proform alternation.

	Feature	Importance
1	plus_xposPOS	2.927
2	plus_xposNNS	1.757
3	plus_xposSYM	1.572
4	plus_xposFW	1.571
5	plus_xposEX	1.47
6	plus_xposJJS	1.425
7	plus_xposNN	1.383
8	plus_xposWRB	1.378
9	minus_xposRBR	1.304
10	plus_xposWP	1.281

4.5. L1 and task effects

Since the CELVA.Sp contains only French learners of English, we tested whether the results could be limited to an L1 effect with the proform alternation (Wang et al., 2024). To do so we tested with/without L1 in the CEFR prediction model applied to the EFCAMDAT test set.

We tested this idea on the proform alternation by introducing L1 as a extra predictor variable and its interactions with the three variants. Results, obtained with an ordinal logistic regression model including odds ratios, remained unchanged. For instance, odds ratio IT = .995 (without L1) when we obtain .996 with the L1 included in the model. THAT has an odds ratio of 1.011 for both models. THIS has 0.998 and IT changes to .997. No confounding effect could be linked to the learner's L1, i.e., upon controlling for L1, the effect of our alternation proform probability on CEFR remained the same. In other terms, this suggests that the link between alternations and proficiency is not impacted by the learners' L1s.



(a) Median probabilities of IT. (b) Median probabilities of THAT. (c) Median probabilities of THIS.

Fig. 5. Normalized form frequency across topics of the EFCAMDAT.

One important aspect to consider is that certain tasks in the corpus may elicit specific vocabulary and/or specific syntactic constructions. Previous research using the EFCAMDAT corpus has revealed effects of task type and instruction, with some tasks showing, for example, a higher number of pronouns (Alexopoulou et al., 2017), or a larger amount of complex noun phrases (Michel et al., 2019). Given the attested task effects in the corpus, it seems plausible that some tasks might be inadvertently eliciting a preference for one variant of the alternation. This might happen if one of the alternation variants is repeatedly used in the instructional prompt, making this variant more salient than the others. Since each task is linked to a particular level, task effects might be a confounding factor in our study, because potential task effects affecting the alternation would not appear consistently across the different proficiency levels.

It is worth noting that there are no predefined task categories in the EFCAMDAT corpus. Prior studies have established categories by inspecting the prompts and establishing similarities. For instance, Michel et al. (2019) propose the following task types: argumentation, description, instruction, narrative, comparison, and list/form. A full study of task effects in relation to the alternations is beyond the scope of this paper, but we qualitatively explored part of the data to investigate whether task effects might be a confounding factor in our study.

We chose the syntactic alternations formed by IT-THAT-THIS as an example, and plotted the normalized form frequency across topics of the EFCAMDAT by CEFR levels, which is shown in Fig. 5. The distribution of THAT across CEFR shows a correlation with topic. This is an interesting case because in our model, the probability of THAT increases with higher proficiency. However, Fig. 5 shows some high-probability outliers at the A1 level, which is not expected. These are defined as values of normalized frequency greater than 10. We examined the *EnglishTown* tasks represented by these outliers to analyze how much of our results could be explained by the instructional prompts. Out of the 14 outliers, 9 correspond to the same task, entitled *Taking inventory in the office* (Topic ID 2), in which students are asked to list the furniture and objects located in an office. While the use of THAT is not encouraged by the prompt, students misuse THAT instead of existential THERE when listing the objects ("That are a lot of computers. That are a lot of chars. That are a lot of desk. That are a lot of mouses. That are some flowers" EFCAMDAT writing ID 924754). In the other three tasks (Topic IDs 7, 12, 21), students were asked to describe or give suggestions about clothing, or buy clothes from a catalogue, and were shown pictures of clothing items. It could be that these tasks, by using images, might be eliciting the use of THAT ("That's is nice and cheap. The purple top is very nice, but that's is too expensive", EFCAMDAT writing ID 309259). It is also interesting that these three tasks target the same vocabulary. Overall, it seems that task effects might account for some of the outliers, but these are also related to misuses of the variants of the alternation, particularly at low proficiency levels. In other words, there is no strong evidence that the task impacts the link between the alternation and proficiency.

5. Discussion

At the start of this paper, we raised three research questions in which we inquired about the relationships between alternations and functions. Our experimental set-up was designed to operationalize, extract, predict and evaluate the predictions of syntactic alternations in terms of proficiency. The purpose was to evaluate the use of probabilities of occurrence of syntactic alternation variants as criterial features of proficiency. We first questioned the potential mapping between alternations and meaning, leading to the identification of sets of forms to be chosen from in the same contexts. We mapped seven alternations to specific functions including, proforms, quantifiers, relativizers, multinoun structures and articles. For each of these alternations we operationalized their extractions with the use of consistent queries relying on multiple annotation layers in the corpora. These extractions were evaluated and results showed a very satisfactory performance.

The second question of the operationalization of alternations was central in our study. We used a novel approach to measure possible occurrences of forms. Rather than using proportions, as advocated in Gaillat et al. (2022), we used probabilities reflecting competition within each alternation. Each model outputs a probability vector of its variants and thus operationalizes the concept of competition. The benefit of this approach is that, instead of relying on the actual token used, the model relies on the contextual features that trigger the form. In fact, the model simulates what a learner would be likely to say for a specific slot of the context. Each alternation model functions as an artificial learner specialized in certain grammatical constructions. Compared with the form

counting approach, the probabilist approach reflects the context produced by the learner rather than just the final word choice. It focuses on the syntactic construction choices made around the form. It adapts to how a learner prepares the instantiation of a form. But is this method correlated with proficiency?

The validity of the construct was the final and crucial question of the study. The purpose was to evaluate whether the alternation models' predictions could be linked to proficiency. Results included odds ratios showing the propensity of an alternation to influence proficiency. A number of alternations appear to be significantly associated with proficiency, albeit weakly. This raises the question of their actual importance for proficiency evaluation.

5.1. Explaining CEFR variance with just alternations?

The percentage of CEFR variance explained with our seven alternations requires discussion since the odds ratios obtained in our results were statistically significant, but all close to 1 (suggesting lack of practical significance). We computed[8] pseudo-R^2. This was to analyze how much of CEFR variance was actually captured by our regression models. The R^2 values were low (for the Proform alternation, R^2 was 1.2% of the variation in CEFR). As a point of reference, Crossley and Kyle (2019) reported R^2 of 20% for a model made up of three cohesion variables. Our results showed alternation odds ratios close to 1, but generally consistent across the internal and external test sets. We thus investigated further the practical significance of the alternation construct.

To test this idea we ran a CEFR prediction model (ordinal logistic regression) including all the alternation variants as features with CEFR as the outcome variable in the test set of the EFCAMDAT. In this combined model we found $R^2 = 0.096$, meaning that 9.6% of the variation in CEFR was explained by the alternations probabilities. This value matches the 10th percentile below which Plonsky and Ghanbar (2018) consider to be somewhat small or modest in terms of variance explained. This suggests that combining the alternations together helped to better understand proficiency. The low R^2 value makes sense if one considers that the seven alternations studied only represent a portion of potentially many more syntactic alternations. We also compared the model with a base-line model including *Average Sentence Length* as an independent variable. We obtained 50.3% balanced accuracy. Conversely, the seven alternation metrics combined give 57.1% balanced accuracy. The increase is not very high, but we would like to stress that the seven alternation metrics cannot be expected to explain proficiency as a whole. Many more microsystems of alternations exist and there are other syntagmatic factors that play a role in proficiency including CAF-related indices.

In addition, Table 8 shows that some variants of some alternations are consistent across the two test sets. These variants are more generalizable because, despite differences regarding the diversity of tasks, the differences in L1 and other sociological differences between the two corpora, odds ratios remained similar. Another argument in favor of alternations is that the alternation markers are omnipresent in (learner) texts, and are more likely to account for CEFR variance than other tokens. What is the likelihood that any other word change in a text may trigger a change in the CEFR level? Alternation candidates correspond to low contexts,[9] which indicates their tendency to accept few candidates per paradigmatic slot. As a result, their occurrence is more common than high entropy contexts such as nouns or verbs.

5.2. Causality vs association

While we find evidence for associations between the probability of using an alternation form and CEFR, this does not mean that increasing this probability will necessarily *cause* an increase in CEFR level. Assessing causality is not possible using these observational data. For instance, a learner in the EFCAMDAT wrote "That are a lot of computers. That are a lot of chars. That are a lot of desk. That are a lot of mouses. That are some flowers". This example shows that using THAT does not necessarily make a writing better. In other words it, is not causal. This is important because even if an alternation showed association, it would be risky to claim that students should write more of a particular variant than the others in order to improve their level. Cross-examination with proficiency should be conducted in order to narrow down potential causality. Nevertheless, association is a first indicator of a potential issue. The advantage of the alternation construct is that it points to a grammatical construct that can be more easily interpreted than complex holistic ratios such as the Type Token Ratio (TTR). Associating an interpretable variable to proficiency helps understanding what makes a learner writing better or worse. This can be very helpful within the context of a Computer-Assisted Language Learning system focused on explaining proficiency classification.

5.3. Designing learner trajectories

Trajectory representation is a matter of using CEFR as an operationalization of interlanguage stages. In this respect, syntactic alternation probability distribution plots provide fine-grained views of the gradual changes, albeit subtle, that can be observed for each syntactic alternation. Fig. 3 illustrates this with the profoms, in which we see a small increase in the probability of use of THAT as CEFR increases. The syntactic alternation construct can be seen as a method to measure and visualize learner trajectories with meaningful form-function mappings. This approach is similar to that of Biber and Staples's (2011) in which single form-function

[8] $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$, i.e., we calculate the R^2 as 1 minus the ratio of the sum of squared errors explained by the model to the total sum of squares.

[9] This can be tested with any Large Language Model, for example with the Hugging Face interface for BERT https://huggingface.co/google-bert/bert-base-uncased. A sentence like "If we know how to use it, [MASK] will improve our way of life" will output probabilities of occurrences much higher for the members of the proform alternation: *it* (0.968), *this* (0.012), *that* (0.010) vs. *we* (0.003) or *they* (0.002).

mappings are analyzed as a function of proficiency levels and registers across the BAW corpus (Staples et al., 2022). Staples et al. examined the development of complexity features across university levels. Overall, they found that L1 writers used more phrasal features and fewer clausal features. Their results illustrated the learning trajectories, showing the variations in frequencies per feature type. O'Keeffe et al.'s project (2017) provided a mapping of lexico-functional patterns to CEFR levels via the use of multiple criteria including frequency, range of users and accuracy thresholds. In doing so, they created a map of patterns as a function of their onset across levels. The computations of patterns were based on occurrences.

In Saussure's representation of language, language (*langue*) is a system of systems. In our examples of syntactic alternations, we focused on alternations related to nominal structures and their determinations. Could we represent learner trajectories with such a limited linguistic scope? We acknowledge that the list of syntactic alternations that we implemented is not exhaustive. We have started developing methods with other forms. Among the candidates that we consider are modals, but the UD annotation scheme does not allow queries for the epistemic/root distinction of modals. Defining new syntactic alternations relies on the ability to design extraction queries that capture all elements of an syntactic alternation. Some alternations may be very semantic which makes their extraction more difficult. As more tools are being developed for the characterization of semantic and pragmatic features, it might become easier to extract the related syntactic alternation.

6. Conclusion and perspectives

In this paper we reported on the findings of a study about the link between syntactic L2 proficiency and syntactic alternations, i.e., groups of forms in paradigmatic competition when learners make their choices of words for specific linguistic functions. Instead of relying on frequency counts of actual words, we designed metrics indicating the probability of a form vs its competitors, hence quantifying the likelihood of a learner to use one of the forms depending on the context. Our purpose was to evaluate the validity of alternations with regard to proficiency.

We adopted a machine-learning approach which relied on the EFCAMDAT dataset to train seven alternation models. These models were subsequently tested on an internal and an external test set. As the data were annotated with CEFR levels, we evaluated the associations between syntactic alternation probabilities and CEFR levels. Results showed that all syntactic alternations were significant but yielded low impact if taken individually. However, their influence was shown to be impactful if taken as a group. These alternations and their measurement method suggest that it is possible to use them as part of broader-purpose CALL systems focused on proficiency assessment and linguistically grounded explanations.

Alternation models could be exploited in L2 AES systems for proficiency prediction. As they capture specific form-function mappings, their output probability scores could be used to support specific linguistic feedback. Based on internal feature-importance mechanisms, an AES could identify some alternation metrics as being associated with specific proficiency levels. The AES could then identify which of the variants are likely to improve proficiency, guiding learners towards better use of a variant.

Our approach appears to be precursory to language models in that it uses contexts for form predictions, but it focuses on restricted sets of linguistic forms. Language models provide probabilities based on the entire vocabulary set of their training corpus. Using large language models for the analysis of alternations would be a logical next step. Their predictive power may provide finer results and pave the way towards the creation of models simulating artificial learners.

CRediT authorship contribution statement

Cyriel Mallart: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Andrew Simpkin:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Nicolas Ballier:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Paula Lissón:** Writing – review & editing, Writing – original draft. **Rémi Venant:** Writing – review & editing, Software, Methodology. **Bernardo Stearns:** Software, Data curation. **Jen-Yu Li:** Software, Data curation. **Thomas Gaillat:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by Agence Nationale de la Recherche (ANR), the French national research agency, as part of the Analytics for Language Learning[10] (A4LL) project. Grant number ANR-22-CE38-0015-01.

[10] See https://lidile.recherche.univ-rennes2.fr/en/article/a4ll-analytics-language-learning.

Table A.10
Features used for the article alternation model.

Feature type	Feature description
POS	Left context 5-gram POS (UD and Penn Tree Bank (PTB))
POS	Right context 5-gram POS (UD and PTB)
POS	Head POS (UD and PTB)
Dependency	Head-dependency relation between form and head
Dependency	Normalized dependency distance to root
Tokens	Head token's position in sentence
Tokens	Position of syntactic alternation token
Morphology	Number of head
Nationality	Nationality of the learner

Table A.11
Features used for the duration alternation model.

Feature type	Feature description
POS	Left context 5-gram POS (UD and PTB)
POS	Right context 5-gram POS (UD and PTB)
Dependency	Head-dependency relation between form and head
Dependency	Normalized dependency distance to root
Tokens	Position of syntactic alternation token
Morphology	token number in 2-gram right context and in 1-gram left context
Nationality	Nationality declared by learner

Table A.12
Features used for the quantifier any/some alternation model.

Feature type	Feature description
POS	Left context 5-gram POS (UD and PTB)
POS	Right context 5-gram POS (UD and PTB)
POS	Head POS (UD and PTB)
Dependency	Head-dependency relation between form and head
Dependency	Normalized dependency distance to root
Tokens	Head token's position in sentence
Tokens	Position of syntactic alternation token
Morphology	token number in 2-gram right context and in 1-gram left context
Nationality	Nationality of the learner

Table A.13
Features used for the quantifier many/much alternation model.

Feature type	Feature description
POS	Left context 5-gram POS (UD and PTB)
POS	Right context 5-gram POS (UD and PTB)
POS	Head POS (UD and PTB)
Dependency	Head-dependency relation between form and head
Dependency	Normalized dependency distance to root
Tokens	Head token's position in sentence
Tokens	Position of syntactic alternation token
Morphology	Token number in 1-gram right context
Nationality	Nationality of the learner

Appendix A. Tables of selected features for each alternation prediction model

See Tables A.10–A.15.

Appendix B. Annotation guidelines for the creation of the Gold Standard including alternation forms

Annotators were given a spreadsheet including one observation per line, i.e., an syntactic alternation form. They were required to select the correct form among a list of possible candidates. The observations included proforms but also irrelevant forms used as disturbing variables.

Annotation manual for alternation patterns

1. Open file with annotator's initials
2. Use the column annotation
3. Select an annotation cell
4. Read the sentence next to the cell and identify the place holder for the token to annotate. It is between two stars e.g., *this*

Table A.14
Features used for the multinoun alternation model.

Feature type	Feature description
POS	Left context 5-gram POS (UD and PTB)
POS	Right context 5-gram POS (UD and PTB)
POS	UPOS of the head of the dependency relation
Dependency	Head-dependency relation between form and head
Dependency	Normalized dependency distance to root
Nationality	Nationality of the learner

Table A.15
Features used for the relativizer alternation model.

Feature type	Feature description
POS	Left context 5-gram POS (UD and PTB)
POS	Right context 5-gram POS (UD and PTB)
POS	Head's POS (UD and PTB)
Dependency	Head-dependency relation between form and head
Dependency	Normalized dependency distance to root
Tokens	Head token's position in sentence
Tokens	Position of syntactic alternation token
Morphology	Mood of token in 1-gram right context
Morphology	Verb of token in 1-gram right context
Morphology	Tense of verb token in 1-gram right context
Morphology	Tense of head if verb
Morphology	Number in 1-gram left context
Nationality	Nationality of the learner

Table B.16
Description of alternation microsystems and their variants.

Syntactic alternations	Description	
Quantifier 1		
	any	as a determiner
	some	as a determiner not as an adverbial
Articles		
	A/an	Article "A" as a determiner
	THE	Article "THE" as a determiner
	Article 0	Nouns without any determiner. As a proxy we list *nouns* that have neither determiner nor possessive pronoun dependency relation. In case there is a THE or A article in front of that noun, select the value corresponding to that article. If it is introduced by a quantifier (fewer, many, any...), select none.
Proforms		
	IT	It as an proform only, not extrapositional e.g. "it's ridiculous that they've given the job to PAt", nor impersonal e.g. "It seemed that/as if things would never get any better.". it-cleft constructions, e.g."It was your father who was driving - No it wasn't not, it was me." or weather/time it e.g. "It's only two weeks since she left." "It's raining."
	THIS	only as proform, not as determiner, nor adverbial
	THAT	only as proform, not as determiner, nor adverbial, nor relativizer nor complementizer.
Multinoun		For the multinoun syntactic alternations, the *last* word of the pattern is between two stars *. For instance: The university *car*; The university's *car*; The car of the *university*
	N of N	Any time a noun appears in a N of N construction
	NN	In cases of NN it can be either first or second position. e.g "I am studying materials science in an *engineering* school". Here consider that the target to evaluate is Engineering school even if it is the first N that is between stars. NOTE: this pattern does NOT include ADJ + NN of course.
	N's N	Any time a noun appears in a N's N construction
Duration syntactic alternations		
	FOR	"For" used to express a lasting period of time (translates as "pendant" in FRench). Not to be confused with expression of purpose. e.g. "I want to do this for a gap year." or reason e.g. "thanks for doing xyz"

(*continued on next page*)

5. When selecting a cell, press "alt" key and arrow-down key to see the possible values to choose from. The "none"; value means the pattern does not correspond (because it has a different function in the context or because, for evaluation purposes, we have taken sentences that do not include the patterns).

Note: for ease of use and speed, it is advisable to use the keyboard keys.

Table B.16 lists the patterns and the definitions to comply with:

Table B.16 (*continued*).

Syntactic alternations	Description	
	SINCE	"Since" used as a point of departure in time
	DURING	"During" used for the expression of a lasting period of time
Quantification		
	MUCH	Used to express quantity
	MANY	used to express quantity
Relativizers		
	THAT	Uses of "that" as relative pronoun only, NOT as proform, determiner, complementizer or adverbial.
	WHICH	Uses of "which" as relative pronoun only, not as interrogative. NOTE: Watch relative pronouns as objects of verb.
	WHO	Uses of "who" as relative pronoun only, not as interrogative. NOTE: be careful with cases where WHO has no apparent antecedent: A who relative clause introduced by verb, e.g. "You can meet who you like" (Larreya & Rivière, 1991)

Table C.17
Annotation differences for the proform alternation in the Gold Standard.

Writing ID	Sentence to annotate	Annotation 1	Annotation 2	Consolidation
5952	Music makes me going through so much emotions and I love it *:* sadness, happiness, nostalgia.	prf it	none	none
1426	My opinion in the invention on the web is *that* is allowed at the time to start to communicate more easily, to exchange document.	prf that	none	none
7552	And *this* is it, at this very moment it stroke me.	prf this	prf that	prf this
3129	That 's why, I 'm contact you *,* because I need your help for my project.	prf that	none	none
3679	I talked with people, teachers and students who have been or were still there, and I create in me a motivation about *this* project.	prf this	none	prf this
5846	Except *that* is also important to consider the negative outcomes we can get from it.	prf that	none	none

Table C.18
Annotation differences for the quantifier some/any alternation in the Gold Standard.

Writing ID	Sentence to annotate	Annotation 1	Annotation 2	Consolidation
814	On social media, people shows what they think is nice to see, all the good in life but it transforms in bad because we do not see the reality in *any* of these photos.	quant any	none	none

Table C.19
Annotation differences for the quantifier many/much alternation in the Gold Standard.

Writing ID	Sentence to annotate	Annotation 1	Annotation 2	Consolidation
6172	The information is not filtering, he *many* have shocking photos or videos.	quant many	none	none
8313	Today we have *many* people than dislike the vaccine, and they don't make it to their children.	quant many	quant much	quant many
10757	The computer accumulate too *much* heat and the component melt.	quant many	quant much	quant much
10950	In world of mobile application development we have *many* tools which help to make an application.	quant much	quant many	quant many

Appendix C. Differences in GS annotations between the two annotators and consolidation decisions

See Tables C.17–C.22.

Table C.20
Annotation differences for the quantifier multinoun alternation in the Gold Standard.

Writing ID	Sentence to annotate	Annotation 1	Annotation 2	Consolidation
5	Moreover this poem shows the *signification* of statue which is liberty, freedom, integration, american dream.	multinoun n2 n1	none	multinoun n2 of n1
10	One of my favorite game is Outer Wilds, an exploration *game* where you play an archaeologist and an astronaut who travels across a tiny solar system in a spaceship to find clues about an antic civilization.	multinoun n2 n1	multinoun n2 of n1	multinoun n2 n1
73	I have the project to work in the *field* of cybersecurity.	none	multinoun n2 of n1	multinoun n2 of n1
100	Alex Dupont 2 : I choose the science of *education* for many reasons but not necessarily to become school teacher.	none	multinoun n2 of n1	multinoun n2 of n1

Table C.21
Annotation differences for the relativizer alternation in the Gold Standard.

Writing ID	Sentence to annotate	Annotation 1	Annotation 2	Consolidation
No difference				

Table C.22
Annotation differences for the duration alternation in the Gold Standard.

Writing ID	Sentence to annotate	Annotation 1	Annotation 2	Consolidation
4605	Hello, my name is Alex Dupont and i am sending you this letter to explain to you my project *for* the next year.	none	dur for	none
2654	Moreover, this work experience is more importe *for* this year, it 's obligatory for a final evaluation.	none	dur for	dur for
11225	To finish i developed a critical spirit and *since* my redaction his better than before.	dur since	none	none
9012	At the end of my second year of medicine, I worked *for* the first time in an EPHAD (sort of center for elder person) as an auxiliary.	none	dur for	none
8723	That 's why the laser surgery is important, because it gives people an other opportunity to correct their myopia, and this surgery improves your vision *for* life.	none	dur for	dur for
9357	By example, if a person sees a black cat *for* the first time, his eyes see it, send a message to the brain that connect its neurones and creates an engram relative to the black cat.	none	dur for	none
5825	I have a strong passion *for* reading and novels.	none	dur for	none
6285	People no longer have to wait *for* a specific time the news at the TV but they can research everythings at everytime on Internet.	dur for	none	dur for
2337	When the holidays arrived I passed all my time with her to help her to prepare her class *for* the following year.	none	dur for	none
4362	*Since* the first time i saw this Alex Dupont do his job with love and passion, i knew what i would like to do in my studies for the future.	dur during	dur since	dur since

(*continued on next page*)

Appendix D. Quality of syntactic alternation extractions in the GS dataset

See Tables D.23–D.28.

Table C.22 (*continued*).

Writing ID	Sentence to annotate	Annotation 1	Annotation 2	Consolidation
4204	This project will be *for* the next year, and I want to go 6 months.	none	dur for	none
4180	Nevertheless, this remain my plan B *since* my plan	dur since	none	dur since
9498	I think, *for* the while, we can't stop the production of nuclear energy because we didn't find a energy enough efficient to substitute the nuclear energy.	none	dur for	dur for
2029	*Since* 3 years, I am interested by children.	dur during	dur since	dur since

Table D.23
Quality of proform syntactic alternation extractions in the GS.

	precision	recall	f1-score	support
NONE	0.64	0.81	0.72	36
PRF IT	1.00	0.78	0.88	51
PRF THAT	0.88	0.92	0.90	38
PRF THIS	0.95	0.95	0.95	40
accuracy			0.86	165
macro avg	0.87	0.87	0.86	165
weighted avg	0.88	0.86	0.86	165

Table D.24
Quality of quantifier some/any syntactic alternation extractions in the GS.

	precision	recall	f1-score	support
NONE	0.38	0.69	0.49	16
QUANT ANY	0.97	0.81	0.89	48
QUANT SOME	0.90	0.80	0.85	45
accuracy			0.79	109
macro avg	0.75	0.77	0.74	109
weighted avg	0.86	0.79	0.81	109

Table D.25
Quality of quantifier much/many alternation extractions in the GS.

	precision	recall	f1-score	support
NONE	0.37	1.00	0.54	11
QUANT MANY	1.00	0.82	0.90	49
QUANT MUCH	1.00	0.80	0.89	50
accuracy			0.83	110
macro avg	0.79	0.87	0.77	110
weighted avg	0.94	0.83	0.86	110

Table D.26
Quality of relativizer alternation extractions in the GS.

	precision	recall	f1-score	support
NONE	0.62	0.88	0.73	32
REL THAT	0.90	1.00	0.95	36
REL WHICH	1.00	0.85	0.92	47
REL WHO	1.00	0.80	0.89	50
accuracy			0.87	165
macro avg	0.88	0.88	0.87	165
weighted avg	0.90	0.87	0.88	165

Table D.27
Quality of multinoun alternation extractions in the GS.

	precision	recall	f1-score	support
MULTINOUN N2N1	0.66	0.85	0.74	27
MULTINOUN N2OFN1	0.51	0.78	0.62	23
MULTINOUN N2SN1	0.94	1.00	0.97	33
NONE	0.73	0.42	0.54	52
accuracy			0.71	135
macro avg	0.71	0.76	0.72	135
weighted avg	0.73	0.71	0.70	135

Table D.28
Quality of article a/the/zero alternation extractions in the GS.

	precision	recall	f1-score	support
ART A	0.98	0.80	0.88	56
ART NONE	0.61	0.88	0.72	25
ART THE	1.00	0.91	0.95	56
NONE	0.48	0.57	0.52	23
accuracy			0.82	160
macro avg	0.77	0.79	0.77	160
weighted avg	0.86	0.82	0.83	160

References

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (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.

Artstein, R., & Poesio, M. (2008). Inter-coder Agreement for Computational Linguistics. *Computational Linguistics*, *34*(4), 555–596.

Banno, S., Vydana, H. K., Knill, K., & Gales, M. (2024). Can GPT-4 do L2 analytic assessment? In E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. s. Tack, V. Yaneva, & Z. Yuan (Eds.), *Proceedings of the 19th workshop on innovative use of NLP for building educational applications (BEA 2024)* (pp. 149–164). Mexico City, Mexico: Association for Computational Linguistics.

Biber, D., & Gray, B. (2011). Grammatical change in the noun phrase: the influence of written language use. *English Language & Linguistics*, *15*(2), 223–250, Publisher: Cambridge University Press.

Biber, D., Gray, B., Staples, S., & Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. *Journal of English for Academic Purposes*, *46*, Article 100869.

Bresnan, J., Cueni, A., Nikitina, T., & Baayen, R. H. (2007). Predicting the dative alternation. In B. Gerlof, I. Kramer, & J. Swarts (Eds.), *Cognitive foundations of interpretation* (pp. 69–94). Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Bulté, B., & Housen, A. (2012). *Defining and operationalising L2 complexity*. John Benjamins Publishing Company.

Council of Europe (2018). *Common European framework of reference for languages: learning, teaching, assessment: Companion volume with new descriptors*. Strasbourg: Council of Europe.

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, *51*(1), 14–27.

Cushing Weigle, S. (2013). English language learners and automated scoring of essays: Critical considerations. In *Automated assessment of writing: Assessing Writing*, In *Automated assessment of writing: 18*(1).85–99,

de Marneffe, M.-C., Manning, C. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, *47*(2), 255–308, Place: Cambridge, MA Publisher: MIT Press.

DeVore, S., & Kyle, K. (2023). Assessing syntactic and lexicogrammatical use in second language Mandarin writing samples. *Journal of Second Language Writing*, *60*, Article 101014.

Dubois, T., Paquot, M., & Szmrecsanyi, B. (2023). Alternation phenomena and language proficiency: the genitive alternation in the spoken language of EFL learners. *Corpus Linguistics and Linguistic Theory*, *19*(3).

Eguchi, M., & Kyle, K. (2023). L2 collocation profiles and their relationship with vocabulary proficiency: A learner corpus approach. *Journal of Second Language Writing*, *60*, Article 100975.

Ellis, R. (1994). *The study of second language acquisition*. Oxford, United Kingdom: Oxford University Press.

Fleiss, J. L., Levin, B., & Cho Paik, M. (2003). The measurement of interrater agreement. In *Statistical methods for rates and proportions* (pp. 598–626). John Wiley & Sons, Ltd.

Gaillat, T., Simpkin, A., Ballier, N., Stearns, B., Sousa, A., Bouyé, M., & Zarrouk, M. (2022). Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach. *ReCALL*, *34*(2), 130–146, Publisher: Cambridge University Press.

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge open language database (EFCamDat). In R. T. Miller, K. I. Martin, C. M. Eddington, A. Henery, N. Miguel, A. Tseng, A. Tuninetti, & D. Walter (Eds.), *Proceedings of the 31st second language research forum*. Carnegie Mellon: Cascadilla Press.

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching*, *52*(3), 229–252, Publisher: De Gruyter Mouton.

Gries, S. T. (2017). Syntactic alternation research: Taking stock and some suggestions for the future. *Belgian Journal of Linguistics*, *31*(1), 8–29, Publisher: John Benjamins.

Gries, S. T., & Deshors, S. C. (2014). Using regressions to explore deviations between corpus data and a standard/target: two suggestions. *Corpora*, *9*(1), 109–136.

Gries, S. T., & Wulff, S. (2013). The genitive alternation in Chinese and German ESL learners: Towards a multifactorial notion of context in learner corpus research. *International Journal of Corpus Linguistics*, *18*(3), 327–356, Publisher: John Benjamins Publishing Company.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An introduction to functional grammar* (3rd ed.). London: Hodder Arnold.

Hawkins, J. A., & Filipović, L. (2012). Criterial features in L2 English: specifying the reference levels of the common European framework. Cambridge: Cambridge University Press.

Jäschke, K., & Plag, I. (2016). The dative alternation in German-English interlanguage. *Studies in Second Language Acquisition*, *38*(3), 485–521.

Kinne, A. (2020). Particle placement in English L1 and L2 academic writing - presse. Louvain-La-Neuve, Belgique: Presses universitaires de Louvain.
Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? In *Studies in writing assessment in new zealand and Australia*: *Assessing Writing*, In *Studies in writing assessment in new zealand and Australia*: *vol. 16*(2).81–96,
Kuhn, M. (2015). caret: Classification and regression training. _eprint: 1505.003.
Kuiken, F. (2023). Linguistic complexity in second language acquisition. *Linguistics Vanguard*, *9*(s1), 83–93, Publisher: De Gruyter Mouton.
Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication*. Dissertation, Georgia: Georgia State University.
Kyle, K., Crossley, S., & Verspoor, M. (2021). Measuring longitudinal writing development using indices of syntactic complexity and sophistication. *Studies in Second Language Acquisition*, 1–32, Publisher: Cambridge University Press.
Labov, W. (1972). Sociolinguistic patterns. *Sociolinguistic patterns*, England: U. Pennsylvania Press, Oxford, Pages.
Larsson, T., Paquot, M., & Plonsky, L. D. (2020). Inter-rater reliability in learner corpus research: Insights from a collaborative study on adverb placement.
MacWhinney, B., Bates, E., & Kliegl, R. (1984). Cue validity and sentence interpretation in English, German, and Italian. *Journal of Verbal Learning and Verbal Behavior*, *23*(2), 127–150.
Mallart, C., Simpkin, A., Venant, R., Ballier, N., Stearns, B., Li, J. Y., & Gaillat, T. (2023). A new learner language data set for the study of English for specific purposes at university level. *vol. 1*, In *Proceedings of the 4th conference on language, data and knowledge - LDK 2023* (pp. 281–287). Vienna, Austria.
Michel, M., Murakami, A., Alexopoulou, T., & Meurers, D. (2019). Effects of task type on morphosyntactic complexity across proficiency: Evidence from a large learner corpus of A1 to C2 writings. *Journal of Instructed Second Language Acquisition*, *3*(2), 124–152.
Mizumoto, A. (2023). Calculating the relative importance of multiple regression predictor variables using dominance analysis and random forests. *Language Learning*, *73*(1), 161–196, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/lang.12518.
Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), Article 100050.
O'Keeffe, A., & Mark, G. (2017). The English grammar profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics*, *22*(4), 457–489, Publisher: John Benjamins.
Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education / Internationale Zeitschrift Für Erziehungswissenschaft / Revue Internationale de L'Education*, *14*(2), 210–225.
Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145, Publisher: SAGE Publications Ltd.
Paquot, M., Grafmiller, J., & Szmrecsanyi, B. (2019). Particle placement alternation in EFL learner vs. L1 speech: assessing the similarity of probabilistic grammars. In A. Abel, A. Glaznieks, V. Lyding, & L. Nicolas (Eds.), *Widening the scope of learner corpus research - presses universit* (pp. 71–92). Louvain-La-Neuve, Belgique: Presses Universitaires de Louvain.
Picoral, A., & Carvalho, A. M. (2020). The acquisition of Preposition + Article contractions in L3 Portuguese among different L1-speaking learners: A variationist approach. *Languages*, *5*(4), 45, Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.
Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *The Modern Language Journal*, *102*(4), 713–731, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/modl.12509.
Safari, F., & Ahmadi, A. (2023). Developing and evaluating an empirically-based diagnostic checklist for assessing second language integrated writing. *Journal of Second Language Writing*, *60*, Article 101007.
Shatz, I. (2020). Refining and modifying the EFCAMDAT: Lessons from creating a new corpus from an existing large-scale English learner language database. *International Journal of Learner Corpus Research*, *6*(2), 220–236, Publisher: John Benjamins.
Staples, S., Egbert, J., Biber, D., & Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. *Written Communication*, *33*(2), 149–183, Publisher: SAGE Publications Inc.
Staples, S., Gray, B., Biber, D., & Egbert, J. (2022). Writing Trajectories of Grammatical Complexity at the University: Comparing L1 and L2 English writers in BAWE. *Applied Linguistics*, amac047.
Straka, M., Hajič, J., & Straková, J. (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* (pp. 4290–4297). Portorož, Slovenia: European Language Resources Association (ELRA).
Vajjala, S. (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.
Wang, H., Wang, G., Wang, N., & Wang, L. (2024). Effects of speaker types and L1 backgrounds on the linguistic complexity of learners' writing. *International Journal of Applied Linguistics*, *34*(2), 692–708, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ijal.12526.
Wulff, S., & Ellis, N. C. (2018). Usage-based approaches to second language acquisition. In *Studies in bilingualism*, *Bilingual cognition and language: the state of the science across its subfields* (pp. 37–56). Amsterdam, Netherlands: John Benjamins Publishing Company.
Wulff, S., & Gries, S. T. (2015). Prenominal adjective order preferences in Chinese and German L2 English: A multifactorial corpus study. *Linguistic Approaches To Bilingualism*, *5*(1), 122–150, Publisher: John Benjamins.
Wulff, S., Lester, N., & Martinez-Garcia, M. T. (2014). That-variation in German and Spanish L2 English. *Language and Cognition*, *6*(2), 271–299.
Yamashita, T. (2024). An application of many-facet Rasch measurement to evaluate automated essay scoring: A case of ChatGPT-4.0. *Research Methods in Applied Linguistics*, *3*(3), Article 100133.
Yancey, K. P., Laflair, G., Verardi, A., & Burstein, J. (2023). Rating Short L2 Essays on the CEFR Scale with GPT-4. In E. Kochmar, J. Burstein, A. Horbach, R. Laarmann-Quante, N. Madnani, A. s. Tack, V. Yaneva, Z. Yuan, & T. Zesch (Eds.), *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)* (pp. 576–584). Toronto, Canada: Association for Computational Linguistics.