Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach

Thomas Gaillat, Andrew Simpkin, Nicolas Ballier, Bernardo Stearns, Annanda Sousa, Manon Bouyé, Manel Zarrouk

Predicting CEFR levels in learners of English: the use of microsystem criterial features in a machine learning approach

Thomas Gaillat
Université Rennes 2, France (thomas.gaillat@univ-rennes2.fr)
Andrew Simpkin
School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway (andrew.simpkin@insight-centre.org)
Nicolas Ballier
Université de Paris, France (nicolas.ballier@univ-paris.fr)
Bernardo Stearns
Data Science Institute (DSI) National University of Ireland, Galway (bernardo.stearns@insight-centre.org)
Annanda Sousa
Data Science Institute (DSI) National University of Ireland, Galway (annanda.sousa@insight-centre.org)
Manon Bouyé
Université de Paris, France (manon.bouye@etu.u-paris.fr)
Manel Zarrouk
Université Sorbonne Paris Nord, France (zarrouk@lipn.univ-paris13.fr)

Abstract
This paper focuses on automatically assessing language proficiency levels according to linguistic complexity in learner English. We implement a supervised learning approach as part of an Automatic Essay Scoring system. The objective is to uncover Common European Framework of Reference (CEFR) criterial features in writings by learners of English as a foreign language. Our method relies on the concept of microsystems with features related to learner-specific linguistic systems in which several forms operate paradigmatically. Results on internal data show that different microsystems help classify writings from A1 to C2 levels (82% balanced accuracy). Overall results on external data show that a combination of lexical, syntactic, cohesive and accuracy features yields the most efficient classification across several corpora (59.2% balanced accuracy).

Keywords: microsystem; criterial features; supervised learning; language functions; Automatic Essay Scoring; linguistic complexity

1. Introduction

Proficiency assessments are an essential requirement for language education centres both at individual and institutional levels. For individuals, learning a language requires regular assessments so that learners and teachers can focus on specific areas to train upon. For institutions, there is a growing demand to group learners homogeneously in order to set

adequate teaching objectives and methods. The design and organisation of language assessment tests are labour-intensive and thus costly. In this context, automatic essay assessment may appear as a solution.

Automating assessment is conducted with Automatic Essay Scoring systems (AES). Initially grounded in rule-based approaches (Page, 1968), more modern systems rely on probabilistic models based on Natural Language Processing (NLP) tools exploiting learner corpora (Meurers, 2015). Some of these models depend on the identification of linguistic features used as predictors of writing quality. In L2 studies, features belong to three dimensions, i.e. Complexity, Accuracy and Fluency (CAF) (Housen et al., 2012; Ortega, 2009; Wolfe-Quintero et al., 1998). Some of these features operationalise complexity and act as criterial features in L2 language (Hawkins & Filipović, 2012). They help build computer models for error detection and automated assessment and, by using model explanation procedures, their significance and effect can be measured. Recent work on identifying criterial features has been fruitful, as many studies have addressed many types of features. However, to the best of our knowledge, few studies have tried to test features of several dimensions within a single model (Tack et al., 2017; Volodina et al., 2016) to investigate how they compare.

In addition, many of the developed models use features that quantify text items on the syntagmatic axis. For instance, the type-token ratio computes the number of tokens in relation to other elements of the syntagmatic chain. This approach relies on categorising linguistic forms distinctly without relating them to possible substitutes in the same position and with the same language function, thus ignoring the relationships that exist between forms on the paradigmatic axis. The way learners select forms of a specific function is not captured in current feature collection methods. Form variations of a given linguistic function (Ellis, 1994) need to be accounted for and a solution may be found in operationalising the notion of microsystem (Gentilhomme, 1979; Py, 1996).

Our proposal is to use a machine learning approach to test criterial features of many dimensions within a single model. The purpose is to provide answers on their respective importance. We also test new functional features that capture functional variations within single linguistic microsystems.

2. Theoretical background

2.1 A multidimensional set of 'criterial features'

Initiated with the Threshold project (Ek & Trim, 1998) and increasingly active in recent years, research on criterial features has focused on linking linguistic properties to L2 proficiency and to the levels of the Common European Framework of Reference for languages (CEFR). However, since the CEFR descriptors used by examiners are not explicitly linked to any linguistic properties at any of the six levels, the research on criterial features aims at identifying these properties (Hawkins & Buttery, 2010).

Among the three components of L2, complexity includes *absolute*, *linguistic* complexity which focuses on quantitative features, i.e. "the number of discrete components that a language feature or a language system consists of, and as the number of connections between the different components" (Housen et al., 2012, p. 24). The two authors further divide linguistic complexity into *system* and *structure* complexity.

There are two main approaches in the identification of criterial linguistic features for proficiency. The first one falls into the *structure* category endorsed by projects like the English Profile project (O'Keeffe & Mark, 2017) or the Global Scale of English project (De Jong & Benigno, 2017). Relying on quantitative methods applied to learner corpora (including errors), specific grammatical or lexical forms and syntactic patterns have been

mapped to specific CEFR levels, forming the original definition of criterial features. The second approach falls into the *systemic* category of complexity as it focuses on the learners' L2 system as a whole. It relies on global measurements in texts and provides information on the range, size, and variety of different forms and structures. The literature abounds with such metrics, starting with the ubiquitous Type Token Ratio (TTR). With the advent of computational methods applied to learner corpora (Granger et al., 2007), many types of *system* complexity metrics have been put to the test as criterial features.

The first group of metrics includes lexical complexity metrics. These measures are based on word counts, lexicons and reference corpora. They were tested as predictive features of learner levels in terms of usage and properties (Crossley et al. 2011; Lu 2012).

The second group of measures corresponds to syntactic complexity. By applying pattern extraction, phrases of different types are detected and counted, giving insight in terms of properties and usage (Lu 2010; Chen & Zechner, 2011; Khushik & Huhta, 2019; Lan et al., 2019). The results of the research showed that correlations exist between CEFR levels and certain features (Lu, 2010, 2014).

Semantic and pragmatic features were also tested in studies including cohesion (Crossley et al., 2016; Crossley & McNamara, 2012) and semantic measurements based on reference corpora (Kyle & Crossley, 2014). Errors, or negative properties of interlanguage, were also tested. Ballier et al., (2019) showed that error-tag frequencies could be used as potential proficiency predictors.

As studies became more elaborate, the question of the relative importance of features of all dimensions was raised. Some tools have been developed for the creation of complexity metrics datasets of various dimensions (Chen & Meurers, 2016). Syntactic and lexical complexity metrics were combined (Arnold et al., 2018; Ballier & Gaillat, 2016) as well as semantic measures (Venant & D'Aquin, 2019). Some experimental designs also combined syntactic, lexical, discourse and error features in the form of metrics (Vajjala, 2017) or properties such as POS and n-grams (Garner et al., 2019; Yannakoudakis et al., 2011) or edit distance between erroneous segments and their corresponding target hypothesis (Tono, 2013). All these efforts bore their fruits for the research community and learner data challenges (the ACL Building Educational Applications workshop series) helped fostering techniques and modelling beyond the learner corpus research community. For example, a shared task was organised at the CAp18 conference on Artificial Intelligence in France. A dataset including lexical, readability and syntactic complexity metrics was provided to competitors to predict CEFR levels of French L1 writings in English. Competitors added other features such as ngrams and spelling errors to compute their models (Ballier et al., 2020).

The results of all these studies show that, in spite of their benefits, other complexity measures are required for the characterisation of proficiency levels. Since the CEFR adopts a functional approach, a line of investigation might reside in identifying *system* metrics that also inform on specific functional structures as pointed out by Biber (2020) . One way of approaching the issue could be through the notion of microsystems.

2.2 Microsystems in learners

Microsystems are part of the *structure* complexity construct. They tap into functional complexity because they are composed of several constructions grouped according to functional proximity. Microsystems can be defined as families of competing constructions in a single paradigm. First introduced by Gentilhomme (1979) with personal pronouns in native French, the notion was cross-examined with that of Interlanguage (Py, 1980). Py argued that a microsystem makes it possible to view language as an unstable equilibrium. Interlanguage microsystems take several shapes, including that of autonomous sets of elements.

Gentilhomme (1980) describes learner microsystems as unexpected uses of forms which are evidence of systemic acquisitional processes. Learners develop microsystems which are unstable and transitory in nature (Py, 2000). In terms of syntax, it is possible to illustrate this process with the paradigmatic interactions between forms of the same linguistic function but of different semantic implications.

The article microsystem composed of *a*, *the* or Ø ("zero article") can provide a base for illustrating this view. For a description of Ø, see for instance (Depraetere & Langford, 2012). Let examples (1), (2) and (3) contrast the uses of *the* in three samples from the EFCAMDAT corpus (Geertzen et al., 2013).

(1) "Ladies and Gentlemans, My flat was robbed the previous evening. In coming back at my home, I saw that *the* window was broken." (EFCAMDAT writing ID: 2498)
(2) "What do you think about positive discrimination in *the* companies?" (EFCAMDAT writing ID: 569744)
(3) "Why *the* gender's discrimination is still a problem in our society?" (EFCAMDAT writing ID: 579779)

The use of the article might be expected in (1) due to the associative anaphora linking *flat* and *window*. However, *the* is unexpected in (2) and (3) due to misunderstandings of the generic values of *companies* and *gender's discrimination*. In examples (2) and (3), Ø is in paradigmatic competition with *the* (Depraetere & Langford, 2012, pp. 91–93). Learners use articles with variability, which constitutes an unstable microsystem. As learners use forms and constructions to perform certain speech acts linked to specific language functions, microsystems can be seen as an attempt to operationalise systematic form-function variations (Ellis, 1994, p. 135). Evidence of this process has been examined through the use of *it, this* and *that* in Gaillat (2016).

To capture the variability within microsystems, our proposal is to create metrics that measure the importance of each construction in relation to its counterparts within a given text. Single measures could thus encapsulate the internal variations of multi-variable microsystems. This approach would bridge the gap between *structure* and *system* complexity. Microsystem metrics offer an insight into the evolution of linguistic functions at systemic level across categories such as articles, modal auxiliaries, tenses and nouns. We take these grammatical areas to be representative of potential interlanguage grammar rules in construction and analyse written productions through these lenses of microsystems.

To the best of our knowledge, the literature on criterial features does not include heuristics based on microsystems, nor does it report many studies testing many metrics as criterial features of many dimensions. Our approach includes the definition of some microsystems which are used for specific language functions such as determination or the expression of modal possibility. Our experimental design exploits machine learning algorithms to classify learner writings with many types of metrics including specifically-designed microsystem metrics.

Our research aims are (i) to assess many complexity metrics as potential criterial features (Hawkins & Filipović, 2012) and (ii) to investigate the significance of microsystem metrics as criterial features within the broad spectrum of complexity metrics.

3. Methods

3.1 Corpora

The data used for modeling and measuring the correlation between learner levels and microsystems consists of the Spanish and French L1 subsets of the Education First-

Cambridge Open Language Database (EFCAMDAT), an 83-million-word corpus collected and made available by Cambridge University and its partner, the organization *Education First*. This corpus is made up of learner writings in English and rated by humans. It was annotated with metadata such as learner level, nationality but also, for some texts, errors and part-of-speech tagging. The levels which were assigned to learners are based on the levels from EF's online school, *Englishtown*, with ratings ranging from 1 to 16. Learner levels thus had to be mapped onto CEFR levels. Levels 1-3 correspond to the A1 level and level 16 to the C2 level, as indicated in (Geertzen et al., 2013). Data was selected and manipulated independently of the participation of the Cambridge and Education First research teams.

In our study, 49,817 texts written by 8,851 French and Spanish learners were downloaded from the database. This textual data runs across all Englishtown writing topics and CEFR levels. Tables 1 and 2 give the breakdown for each L1.

Table 1. The EFCAMDAT French dataset

	A1	A2	B1	B2	C1	C2
# of tokens	817,228	888,298	887,987	528,880	138,541	13,689
# of types	561,688	581,317	571,193	320,973	80,722	8161
# of writings	17,605	11,584	8,105	3,514	742	76
Median	36	67	98	134	173	170.5

Table 2. The EFCAMDAT Spanish dataset

	A1	A2	B1	B2	C1	C2
# of tokens	125,500	163,668	228,710	185,094	64,534	5,954
# of types	84,334	106,553	144,295	108,942	37,150	3,620
# of writings	2,572	2,066	2,005	1,176	340	32
Median	38	68	103	143	173	167.5

To test the validity of our models on external data, we used the CEFR ASAG corpus (Tack et al., 2017), a collection of short answers to open-ended questions, written by French L1 learners of English and graded with CEFR levels. It consists of 712 texts written by different learners in response to three questions. We used a balanced sample of 299 texts.

3.2 Features

We created new functional metrics based on the notion of microsystems (see Section 2.2). We assume that microsystems are sets of competing constructions (some being more likely for natives, others more prone to be L1-like). Based on intuition, Table 3 provides a list of other potential functional microsystems identified by two expert English teachers and linguists. For instance the nominal microsystem includes three constructions which learners find difficult.

They may use genitive constructions instead of noun+preposition+noun or compound noun constructions. Similarly, other substitutions may be observed among the *can*, *may*, *might*, *could* modals used to express epistemic and radical possibility. Regarding *that*, it has been noticed that confusions occur between the relativizer forms. We also specified a type of error linked to the confusion between the relativizer and complementizer functions.

Table 3. Learner microsystems and their variables.

Microsystems	Function	variables
Nominal constructs	Denomination	determiner genitive; noun-of/for-noun constructions, compound nouns
Modals for possibility	Possibility	*may*; *can*; *might*; *could*
Modals for obligation	Obligation	*must*; *have to*
Proforms	Reference	*it*; *this*; *that*
Articles	Determination	*a*; *the*; *Ø*
Relativisers	Reference	*that; which; who; 0*
Complementizer vs relativizer	Expressing hypotaxis	*that*
Duration/start/date	Expressing time	*For; since; ago; from; during*
Prepositional constructions	Linking entities	*For; to*
Quantifiers	Quantification (Neutral; large; small)	*Some* vs *any; many* vs *much* vs *most; few* vs *little*

Note: Relative pronoun 0 is not included in the operationalisation of the program as the detection of the non-existent tokens remains an obstacle

The microsystems include variability in grammaticality: some of the substitutions among the aforementioned constructions are just semantic differences in the case of modal auxiliaries, others jeopardise grammaticality (*which* versus *who* for animate antecedents). The weighting of the parameters of these different constructions is beyond the scope of this paper.

Finding a method to quantify variability in microsystems at text level could help measuring the importance of specific linguistic functions in L2 systems. To operationalise microsystems, we added a set of metrics relying on paradigmatic relations between forms of similar functions, i.e. microsystem variables as defined in Table 3. For each microsystem xxx (e.g., "modals for possibility"), the frequency of occurrence fff of each variable iii (e.g., "may") in this microsystem was computed within each text jjj (see Eq. 1a). In addition, a

ratio was computed for each variable iii relative to all nnn variables of the microsystem (see Eq. 1b). The absolute and relative microsystem features were computed as follows:

$$MS_A(x_{ij}) = f_{ij} MS_A(x_{ij}) = f_{ij} MS_A(x_{ij}) = f_{ij} \quad (1a)$$

$$MS_R(x_{ij}) = f_{ij} / \sum_{k=1}^{n} f_{kj} \quad (1b)$$

where
x = the microsystem
n = the total number of variables in microsystem x
i = the i-th variable in the set of n variables
j = the j-th text (learner writing)
f_{ij} = the frequency of occurrence of variable i in text j

The microsystem ratios reflect the variations in the proportions of one variable over its paradigmatic competitors. Microsystem features are computed within each writing separately.

The L2 Syntactic Complexity Analyzer (L2SCA) tool (Lu, 2010) was modified in order to capture specific linguistic forms belonging to specific microsystems. The program proceeds in two stages. Firstly, it extracts the constructions used in the microsystems and, secondly, it calculates ratios which operationalise the microsystems. The Tregex module of Stanford CoreNLP (Manning 2014) was used to retrieve constructions including nouns, modal auxiliaries, articles, proforms, relativizers and complementizers. For illustration's sake, we focus on the microsystem of proforms. The Penn Treebank tagset used for the program does not have a specific tag for proforms, so that the *this* proforms were retrieved with the following Tregex patterns:

prf_this1='DT=n1 </[t|T]his/ & >- /NP.*/' (1)
prf_this2='/[t|T]his/ > NN=n1' (2)

Pattern (1) identifies all *this* that are tagged as DT (determiner) and that are the rightmost descendents of Noun Phrase (NP) constituents. Pattern (2) identifies all *this* immediately dominated by a Noun (NN).

The evaluation of the extractions of all the forms specified in microsystems is outside the scope of this paper. Nevertheless, it must be mentioned that most forms are captured with patterns relying on their POS tags (See Appendix 1). It may be argued that evaluating their extraction relates to evaluating POS tagging in learner corpora (accuracy results above 95%). Several papers have established a high level of accuracy in POS tagging learner English (see (Huang et al., 2018; van Rooy & Schafer, 2003). The analysis of proforms is not based on the identification of the tag and previous works support its reliability (Gaillat, 2016, pp. 183–196). The extraction of *this* forms was evaluated by applying distinctive patterns on 2,853 occurrences in the Wall Street Journal subset of the Penn Treebank corpus (Marcus et al., 1993). All *this* proforms were accounted for.

As a result of the extraction process, 51 constructions were incorporated as variables in 29 microsystem metrics (See Appendix 1 for a list of microsystem metrics, their variables and Tregex extraction patterns). The modified version of L2SCA is called L2SCA_microsystem[1]. It also includes the same indices as L2SCA.

In addition to these microsystem features, several other types were extracted and used to compute metrics. The feature types encompass lexical, syntactic, semantic and discourse

1 Available from the project's website: http://www.clillac-arp.univ-paris-diderot.fr/projets/ulysse2019

complexity as well as accuracy. See Appendix 2 for a list of all the implemented metrics and the tools used to compute them. In total, 767 different features were extracted and merged into one dataset to input into the classification models.

3.3 Statistical analysis

There were three aims in this statistical analysis.

1. Test the utility of the novel microsystem features over existing features
2. Compare feature importance
3. Build a prediction model for future learners

We implemented this analysis through a Machine Learning (ML) approach. In principle, an ML analysis relies on observations recorded in a computer model. In our experiments, the observations are made up of the features of the texts linked to their CEFR levels, and their statistical relationships are computed by applying a specific mathematical function, i.e. a model. The model is subsequently used to predict CEFR levels in new observations of features. The analysis performed for each of the three aforementioned aims is summarised below. Analysis (see Code in Appendix 3) was performed using R v3.6 through the {glmnet} (Friedman et al., 2010) and {caret} packages (Kuhn, 2008).

3.3.1 Testing the utility of the microsystem features

In order to test the efficacy of our novel microsystem variables, we built three classification models: (i) using 687 features from previous research as explained in Section 4.2 as a baseline, (ii) adding the 51 microsystem variables introduced in this paper along with 29 microsystem ratios and (iii) adding the 51 microsystem variables introduced in this paper along with 12 interactions (see Appendix 3) involving variables of the same microsystems.

Using dataset (i) we compared multinomial logistic regression, ensemble random forests, linear discriminant analysis, k-nearest neighbours, Gaussian naive Bayes, support vector machine and decision tree classifier. We found the optimal classification model for (i) and applied this model to each set of features (ii) and (iii). We report on the precision, recall, F1-score (F1 = harmonic mean of precision and recall, i.e. $2 \times \frac{precision \times recall}{precision + recall}$), and balanced accuracy (Balanced Accuracy = average of sensitivity and specificity, i.e. $\frac{sensitivity + specificity}{2}$) of each model. Results are presented for each of the six learner classes, and overall by micro-averaging over the classes to take account of different class sizes. Models were run using five-fold cross-validation to allow for testing with multiple random splits of the data. After running these models, results were macro-averaged across cross-validation folds.

Once the model is used to predict learner level in the test set, we perform an error analysis. We define the error group as a three-level categorical variable, i.e. 0 if classification is correct, 1 if classification is one level lower/higher, 2 if classification is two or more levels lower or higher. A one-way analysis of variance is then used to test whether there are mean differences in each feature according to the error group, adjusting for multiple testing across 767 total features, and taking only those p-values of $< 0.05/767$ to be statistically significant.

3.3.2 Comparing microsystem feature importance

A second analysis used multivariable logistic regression, a classifying method for categorical data, to investigate the relative importance of the 51 new microsystem variables and their 29

ratios across learner levels. We split the data based on learner levels (A, B and C) and ran separate logistic regressions on these data using only the microsystem variables. We report on the strongest positive and negative associated features in terms of their Wald test statistic or z-score for each level, i.e. A2 v A1, B2 v B1 and C2 v C1. A positive association suggests the feature is more common in advanced learners, a negative association suggests the feature is less common in advanced learners. We report on the odds ratios of the features to explore how much the use of a feature increases the odds of being an advanced learner.

3.3.3 Building a classification model for future learners

While the optimal model found using all features in 4.5.1 will allow classification of future learners, using over 700 features will also likely overfit to the EFCAMDAT sample data. Therefore we employed a feature selection algorithm, in particular elastic net regression (Zou & Hastie, 2005), which conducts dimension reduction and prediction simultaneously. Elastic Net Regression is a useful classifying method for modelling the relationship between a binary response variable Y and a large number of potential features $X_1, \cdots, X_P$. The regression model used is

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

where $\pi = P(Y=1)$, $\beta_0, \beta_1, \cdots \beta_p$ are regression coefficients and $i = 1, \cdots, n$ observations are available. In cases where the number of predictors P is bigger than n, some form of model selection or dimension reduction is required. Penalized regression is one such tool which shrinks the coefficients $\beta_0, \beta_1, \cdots \beta_p$with several types of penalty available. The elastic net combines two common penalized regression approaches, i) the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1994) and ii) ridge regression (Hoerl & Kennard, 2000). This is useful because the LASSO allows for automatic feature selection by shrinking coefficients of some variables to 0, while the ridge regression penalty excels where features are heavily correlated--which is likely the case for linguistic features.

Five fold cross-validation was used to repeatedly test performance across multiple splits of the data. The performance metrics - precision, recall, F1 score and balanced accuracy were calculated in each fold and summarised using their macro-average (i.e. simply taking the average of the five precision, recall, F1 and balanced accuracy metrics) and standard deviation.

3.4 Data for evaluation

To evaluate the models we applied a twofold strategy. Firstly, we used a subset of the EFCAMDAT dataset as an internal test set and, secondly, we used the CEFR-ASAG external dataset to test the validity of the model and its resistance to overfitting. We used this corpus as it was made up of small writings and was challenging for the dataset on which we had trained our model. The mean of ASAG texts was 157.62 tokens per writing (SD: 81.66) distributed over the six levels, a value typically associated to A1 in our data. Whereas our corpus is heavily biased towards A1, the ASAG corpus has a majority of B1 writings.

The internal test set was sourced randomly from 25% of the EFCAMDAT dataset, resulting in 12,454 texts. Among the seven model types tested the optimal classification performance in the testing dataset was found using multinomial logistic regression.

The external test set was made up of 299 short texts. It was built with the same feature extraction program described in Section 4.3 ran on the CEFR-ASAG corpus texts. Firstly, the optimal classification model from (i) was used to classify with all the features as in (i) and (ii) (see Section 4.4.1). Secondly, following Ockham's razor principle and to avoid overfitting (capturing non-generalizable features), an elastic net method was applied including feature dimensionality reduction.

4. Results and feature analysis

4.1 Testing the utility of the microsystem features

4.1.1 Classification of all six CEFR levels

Among the seven model types tested, the optimal classification performance in the testing dataset was found using multinomial logistic regression. The classifier using previously developed features achieved 80% balanced accuracy. Using the additional microsystem variables, along with their ratios increased performance to 82%, which translated to an extra 249/12,454 writings correctly classified. Full results are given in four tables in Appendix 4. It includes classification performance, the confusion matrix and detailed comparisons with and without microsystem features. One comment about confusions is that they mostly occur with adjacent classes. A closer examination shows that many writings tend to be classified in the lower adjacent class. Note that the appendix only includes one of multiple confusion matrices from Cross-Validation (CV).

We performed an error analysis in those 12,454 test essays - 10,159 were correctly classified, 1,865 misclassified one level higher/lower error, and 430 misclassified two or more levels higher/lower error. From ANOVA, 469 out of 767 features show mean differences between these two groups, indicating which of the features are associated with errors. The top ten of these are shown in Table 3 in Appendix 4.

4.1.2 Comparing microsystem feature importance

The second analysis of the internal testing protocol relied on the logistic regression model and aimed at investigating the relative importance of microsystem variables across the aggregated A, B and C CEFR levels. We measured the impact of microsystem features in each level. There are two types of features. Figures 1, 2 and 3 (available as supplementary material) show features that indicate occurrences of specific variables and others (with the MS prefix) that show microsystems composed of specific variables. The figures show the strongest features of each level in terms of z-score.

Results regarding the A level (Figure 1) reveal four significant microsystems. Nominal constructions (i.e. prepositional, genitive and compound constructions) relative to each other appear to be significant predictors of the A2 level as opposed to the A1 level. The obligation microsystem composed of modals *have to* and *must* also appears as a significant predictor of A2. Likewise, the duration microsystem (based on *for* and *ago*) as well as the quantification microsystem (based on quantifiers *much*, *most* and *many*) both show preference for A2 rather than A1 writings. As the microsystems implement forms of a specific language function, these results may indicate that writings are likely to implement the nominal, obligation, duration and quantification functions as a first step in their progress. Even more so as A1 tasks are mostly with the present tense, so that for/since/ago is probably not tested at this stage.

Results (Figure 2) show that the B level is influenced by two microsystems. The determination microsystem tends to be indicative of the B1 level. The quantification

microsystem with *most* and *many* appears to be indicative of the B1 level too. This trend is to be compared with that of the A level, in which the quantification microsystem is favoured in A2. The level adjacency may indicate that the quantification language function appears and consolidates between A2 and B1 levels. In functional terms, B learners seem to be developing their proficiency by implementing determination and quantification language functions. The B2 level tends to appear as these microsystems stabilise in terms of variable proportions.

For level C writings (Figure 3), the proform microsystem and several specific constructions appear to be significant. The proform microsystem tends to predict C1 as learners overuse *this* compared with *it* and *that*, whereas the microsystem tends to predict C2 as learners increase the relative importance of *that*. This microsystem suggests the onset of anaphoric and deictic reference processes, which corresponds to more complex discourse. With higher discourse complexity, learners tend to increase their use of referential expressions leading to variability in the proform microsystem. The modals *should* and *will* also appear to be significant. This may indicate more elaborate discourse in writing as learners diversify their stance in terms of epistemic or radical modality.

4.2 Building a classification model for future learners

4.2.1 Logistic Regression model for classification using all features

In order to test the validity of the logistic regression model trained on the EFCAMDAT dataset, the same model was used to classify a dataset built from the CEFR-ASAG corpus. Classification according to the 6 CEFR levels showed poor results with 51% balanced accuracy in the ASAG data.

There are several reasons for the loss in balanced accuracy between the two datasets. Firstly, performance in test data randomly taken from the training data is always optimistic, because the test and train sets are very similar. Conversely, the CEFR-ASAG corpus corresponds to shorter contexts and different tasks than the EFCAMDAT corpus. Secondly, the ASAG data have few A1 writings (~16%), while the EFCAMDAT has approximately 40%. This lack of calibration between class populations is not reflected in the model, leading to errors.

4.2.2 Elastic net modelling EFCAMDAT data with feature selection

To limit overfitting and improve classification on external data, we used an elastic net regression model on the EFCAMDAT training set. This method is a classifying algorithm which comes with the benefit of including feature dimensionality reduction, i.e. feature selection. The elastic net model fitted in 178 minutes using a Macbook Pro with 8GB of memory. Using just 44 features classification showed 75.0% balanced accuracy (CI [74.3, 75.8], $p < 0.001$) and 59.2% (CI [53.4, 64.8], $p < 0.001$) on the EFCAMDAT and CEFR-ASAG test sets respectively (see tables in Appendix 4 part B). Compared with the logistic regression model, the elastic net regression model showed lower performance on the EFCAMDAT test set but, most importantly, it improved performance on the CEFR-ASAG test set showing context adaptability.

The elastic net modelling method combines regression with feature selection. In other terms, it employs methods to not only compute best fit for all data points but to also remove non-significant features. In doing so, it combines the smallest set of features for the best classification. In the EFCAMDAT regression model, 44 features are combined. The features

encompass several linguistic dimensions. Table 4 shows how features are distributed according to their linguistic dimensions.

Table 4. Combined features for best model classification with elastic net

Dimension	Features	Tools
Lexical variation	Lexical Variation; Noun Variation; Type Token Ratio	LCA
Lexical sophistication	Brysbaert_Concreteness_Combined_AW COCA_Academic_Bigram_Lemma_Frequency COCA_lemma_academic_bi_MI Component_Brysbaert_Concreteness_Combined_AW Component_Brysbaert_Concreteness_Combined_CW Component_COCA_fiction_tri_2_MI Component_COCA_spoken_Frequency_CW Component_COCA_spoken_tri_2_DP Component_COCA_spoken_tri_2_MI Component_Freq_N_AW Component_MRC_Imageability_AW Component_MRC_Imageability_CW Component_SUBTLEXus_Freq_CW MRC_Concreteness_AW MRC_Familiarity_AW MRC_Imageability_AW eat_tokens_AW	TAALES
Syntactic complexity	MD_OUGHTTO MS_PRF_THAT	L2SCA microsystem
	acad_av_delta_p_verb_cue_type acad_av_faith_verb_cue_type acad_av_lemma_freq acad_av_lemma_freq_type all_av_lemma_freq all_av_lemma_freq_type amod_nsubj_deps_struct conj_and_all_nominal_deps_NN_struct conj_and_nsubj_deps_struct fic_av_lemma_freq fic_av_lemma_freq_type mag_av_lemma_freq mag_av_lemma_freq_type	TAASC

	neg_per_cl news_av_lemma_freq news_av_lemma_freq_type news_lemma_attested poss_nsubj_deps_NN_struct poss_nsubj_deps_struct	
Cohesion	adjacent_overlap_verb_sent conjunctions	TAACO
Accuracy	average_mispelling_every50words	PyEnchant

Among the microsystem features presented in Section 3, the proform microsystem based on *that* appears to be significant when combined with other lexical, syntactic, accuracy and pragmatic features. The modal *ought to*, in its raw frequency, is conjointly significant with the other features. This suggests that sophisticated grammatical markers could be used as criterial features for lexical sophistication.

5. Discussion

The performance in classification of the logistic regression and the elastic net models shows comparable results with those obtained in other studies applying L2 English proficiency classification. To the best of our knowledge, all studies use test sets extracted from the same corpora as their training sets. Likewise, we tested our models internally and best results showed 82% balanced accuracy on the 6-point CEFR scale with a logistic regression model. We even obtained 95% balanced accuracy on a 2 beginner-and-advanced scale, which can be useful for large scale automated groupings of students above and below the B1/B2 border. In comparison, Vajjala (2017) reported 73.2% balanced accuracy on a TOEFL subset categorized according to a 3-point scale. Crossley et al. (2014) reported 55% on another TOEFL subset on a 5-point scale, and Tack et al. (2017) reported 53% balanced accuracy on the ASAG corpus with a 5-point scale.

Error analysis in the confusion matrix of the logistic regression model revealed a substantial number of errors between proficiency levels including non-adjacent class errors. Significant differences are mainly due to errors related to word frequencies and syntactic patterns (Complex Nominals and Verb Phrases). Regarding frequencies, some learners may have written an unexpected number of words for their level. Regarding syntactic patterns, the complex nominal (CN1) feature includes nouns plus adjective, possessive, prepositional phrase, relative clause, participle, or appositive. This broad variety of structures may create noise in the model. For instance, learners of different levels may use the relative clause structure leading to ambiguities in classification.

Compared with the logistic regression presented in this paper, all the aforementioned studies showed the advantage of limiting the number of features and increasing their potential for generalization. Our logistic regression model relies on a large array of features, which makes it prone to overfitting. After reducing dimensionality with the elastic net method presented in this paper, the model classified 75% of the data correctly. This result compares well with the aforementioned performance rates.

In order to measure the potential for generalization of our models, we tested the trained models on external data. The logistic regression model showed signs of overfitting since the

balanced accuracy on external data dropped from 81% to 51%. Conversely, the elastic net model showed a higher ability for generalisation with a 59.2% balanced accuracy on external data. These results show that external validation of models is a necessary step in order to assess the fit of a model and the significance of its features. This appears as an essential step to include in further studies, and it shows the importance of open access to data sources.

In terms of feature significance, our approach was twofold. The first research question was to assess a large array of complexity metrics as potential criterial features. Based on a dataset of 767 metrics and 49,817 observations, an elastic net method helped identify a limited set of significant features. It is important to stress that it is the combination of features that supports the results. In other terms, it would be incorrect to isolate each of the 44 features and give them independent significance. The feature selection showed that it was mostly lexical and syntactic features that supported best classification. These findings are in line with several studies (Crossley et al., 2011; Kyle & Crossley, 2014; Lu, 2014; Vajjala, 2017).

A caveat is in order at this stage. The models were trained mainly on short texts, with a scarcity of data at specific CEFR levels. The models may be sensitive to variations due to differences in instruction tasks implying the use of some microsystems vs. others. Consequently, microsystems and other features may not be captured in sufficient numbers in some classes leading to unclear boundaries between classes.

The second research question was to investigate the significance of new microsystem metrics as criterial features. We tested these features as part of a multinomial logistic regression model. Each microsystem operationalises the paradigmatic relations of competing constructions in learners. The results show that microsystem features contribute to improving CEFR level prediction, albeit to a small extent. The results suggest a series of learning stages. The ratios of nominal constructions relating two nouns, the ratios of modals linked to obligation and the ratios of quantifiers all appear to be indicative of the A level. Concerning the B level, ratios of quantifiers including *most*, *many*, *little* and *few*, as well as ratios including determiners *a, the* and *Ø*, show significance. This suggests that learners introduce quantification between the A2 and B1 levels and that determination starts occurring in significant proportions at B1. The C level shows the proform microsystem as significant as well as specific modals such as *should* and *will.* As discourse complexifies, learners introduce language constructions with higher semantic complexity. Learners construct referential processes by including deictic and anaphoric constructions, and they increasingly take stances as they use deontic and epistemic modality devices. Some features may be subject to task effects e.g., the use of modal *will* in A1 (see Section 4.1.2).

In the context of language teaching, microsystem features might appear very informative. Microsystems contrast forms that compete with each other in the minds of learners. Using them could prove to be fruitful in iCALL systems providing formative feedback based on simple, clear, elaborated manageable units (Shute, 2008). Microsystems are operationalised as simple limited sets of items which are clearly organised according to linguistic functions (Biber et al., 2020). They could be used to build automated feedback on specific language functions as Saricaoglu shows with causal explanations (Saricaoglu, 2019). In addition, the approach could augment the drive towards Data Driven Learning as the system feeds from a corpus to guide learning (Boulton, 2017).

6. Conclusion

In this paper, we have reported a supervised learning approach for the classification of learner writings in English according to the six CEFR proficiency levels. Our hypothesis concerned the use of linguistic metrics in the determination of CEFR levels. Firstly, we assessed the significance of many complexity metrics as potential criterial features in proficiency. The

models show that a combination of lexical, syntactic, accuracy and pragmatic features helps predict CEFR levels. Among all feature types, lexical and syntactic features appear to be very important. In this respect, frequency information extracted from reference corpora favours prediction. Unlike previous research, our study also provides additional external validation with the ASAG corpus. We tested the portability of the models across corpora with different topics and prompts and showed that some features help with model generalisation.

For the second research question, we investigated the significance of newly designed microsystem metrics as criterial features. These metrics are based on learner-specific paradigms including competing constructions. Specific functional features that function paradigmatically have proved to influence the perception of learner writing proficiency by human annotators. Analysis of the results suggests that some microsystems are connected to acquisitional stages operationalised in terms of levels. The study maps specific constructions to levels in functional terms.

Results are also encouraging as part of the development of an AES protype[2]. The project includes an NLP pipeline built upon several state-of-the-art tools measuring lexical, semantic, syntactic, accuracy and pragmatic complexity. The system provides two services: CEFR level prediction and complexity metric extraction. It relies on the Docker technology which makes it deployable as a cloud service (Sousa et al., 2020).

Understanding foreign language acquisition is a long path which involves many dimensions. With experience, language teachers acquire these dimensions intuitively in order to assess and train their students. However, processing students' productions is slow and variable. The research presented here should be seen as a way to invent new tools to assist teachers who would benefit from easy-to-use analytical tools that objectivise the progress of their learners.

7. Acknowledgments

With the financial support of the French Ministry for Europe and Foreign Affairs and the French Ministry of Higher Education, Research and Innovation and the Irish Research Council as part of PHC Hubert Currien Ulysses 2019 (ref 43121RJ).

8. Supplementary material

Note: all appendices and figures are available for download from the IRIS database under the authors' names. See https://www.iris-database.org

9. Ethical statement

The authors declare no conflict of interest regarding the publication and their involvement in other roles. All the documented research was conducted according to the EU's GRDP act. This material is the authors' own original work. It has not been previously published elsewhere, nor is currently being considered for publication elsewhere. The authors present truthful and complete results of their work. All sources used are appropriately cited in the articles. All authors have been involved in substantial work for this paper, and will take public responsibility for its content.

2 See project's website for demo, data set and other resources: http://www.clillac-arp.univ-paris-diderot.fr/projets/ulysse2019

10. References

Arnold, T., Ballier, N., Gaillat, T., & Lissòn, P. (2018). Predicting CEFR levels in learner English on the basis of metrics and full texts. *Proceedings of the 20th Conférence Sur l'Apprentissage Automatique*, 75–82.

Ballier, N., Canu, S., Petitjean, C., Gasso, G., Balhana, C., Alexopoulou, T., & Gaillat, T. (2020). Machine learning for learner English. *International Journal of Learner Corpus Research*, *6*(1), 72–103.

Ballier, N., & Gaillat, T. (2016). Classifying French learners of English with written-based lexical and complexity metrics. *JEP-TALN-RECITAL 2016*, *9*, 1–14.

Ballier, N., Gaillat, T., Simpkin, A., Stearns, B., Bouyé, M., & Zarrouk, M. (2019). A Supervised Learning Model for the Automatic Assessment of Language Levels Based on Learner Errors. In M. Scheffel, J. Broisin, V. Pammer-Schindler, A. Ioannou, & J. Schneider (Eds.), *Transforming Learning with Meaningful Technologies*. Switzerland: Springer International Publishing, 308–320.

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*, 100869.

Boulton, A. (2017). Data-Driven Learning and Language Pedagogy. In Thorne, S. L. & May, S. (eds.), *Language, Education and Technology*. Cham: Springer International Publishing, 181–192). https://doi.org/10.1007/978-3-319-02237-6_15

Chen, M., & Zechner, K. (2011). Computing and Evaluating Syntactic Complexity Features for Automated Scoring of Spontaneous Non-native Speech. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, 722–731.

Chen, X., & Meurers, D. (2016). CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, 113–119.

Crossley, S. A., Kyle, K., Allen, L. K., Guo, L., & McNamara, D. S. (2014). Linguistic Microfeatures to Predict L2 Writing Proficiency: A Case Study in Automated Writing Evaluation. *The Journal of Writing Assessment* 7 (1): 34.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*(4), 1227–1237. https://doi.org/10.3758/s13428-015-0651-7

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. *Journal of Research in Reading*, *35*(2), 115–135. https://doi.org/DOI: 10.1111/j.1467-9817.2010.01449.x

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561–580.

De Jong, J. H., & Benigno, V. (2017). Alignment of the Global Scale of English to other scales: The concordance between PTE Academic, IELTS, and TOEFL. *Pearson: Global Scale of English Research Series*, 18.

Depraetere, I., & Langford, C. (2012). *Advanced English grammar: A linguistic approach*. London: Continuum.

Ek, J. A. van, & Trim, J. L. M. (1998). *Threshold 1990* (Conseil de l'Europe, Ed.). Cambridge: Cambridge University Press.

Ellis, R. (1994). *The Study of Second Language Acquisition*. Oxford: Oxford University Press.

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22.

Gaillat, T. (2016). *Reference in Interlanguage: The case of this and that. From linguistic annotation to corpus interoperability* [Thesis]. Université Paris-Diderot.

Garner, J., Crossley, S., & Kyle, K. (2019). N-gram measures and L2 writing proficiency. *System*, *80*, 176–187. https://doi.org/10.1016/j.system.2018.12.001

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*. Cascadilla Press.

Gentilhomme, Y. (1979). Micro-systèmes linguistiques et langagiers: Fonctions heuristiques et didactiques. Introduction méthodologique. *Travaux du Centre de Recherches Sémiologiques*, *34*, 1–31.

Granger, S., Kraif, O., Ponton, C., Antoniadis, G., & Zampa, V. (2007). Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness1. *ReCALL*, *19*(3), 252–268. https://doi.org/10.1017/S0958344007000237

Hawkins, J. A., & Buttery, P. (2010). Criterial Features in Learner Corpora: Theory and Illustrations. *English Profile Journal*, *1*(01).

Hawkins, J. A., & Filipović, L. (2012). *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework* (Vol. 1). Cambridge University Press.

Hoerl, A. E., & Kennard, R. W. (2000). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *42*(1), 80–86.

Housen, A., Kuiken, F., & Vedder, I. (eds.). (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (Vol. 32). Amsterdam: John Benjamins Publishing Company.

Huang, Y., Murakami, A., Alexopoulou, T., & Korhonen, A.-L. (2018). Dependency parsing of learner English. *International Journal of Corpus Linguistics*, 23(1): 28-54.

Khushik, G. A., & Huhta, A. (2019). Investigating Syntactic Complexity in EFL Learners' Writing across Common European Framework of Reference Levels A1, A2, and B1. *Applied Linguistics*, *amy064*.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, *28*(1), 1–26. https://doi.org/10.18637/jss.v028.i05

Kyle, K., & Crossley, S. A. (2014). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly*. https://doi.org/10.1002/tesq.194

Lan, G., Lucas, K., & Sun, Y. (2019). Does L2 writing proficiency influence noun phrase complexity? A case analysis of argumentative essays written by Chinese students in a first-year composition course. *System*, *85*, 102116. https://doi.org/10.1016/j.system.2019.102116

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496.

Lu, X. (2014). *Computational Methods for Corpus Annotation and Analysis*. Springer.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

Meurers, D. (2015). Learner Corpora and Natural Language Processing. In Granger, S., Gilquin, G. & Meunier, F. (eds.), *The Cambridge Handbook of Learner Corpus*

Research. Cambridge: Cambridge University Press, 537–565.

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. https://doi.org/10.1075/ijcl.14086.oke

Ortega, L. (2009). *Understanding Second Language Acquisition*. Hodder education.

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.

Py, B. (1980). Quelques réflexions sur la notion d'interlangue. *Revue Tranel (Travaux Neuchâtelois de Linguistique)*, *1*, 31–54.

Py, B. (1996). Les données et leur rôle dans l'acquisition d'une langue non maternelle. *Les Carnets du Cediscor. Publication du Centre de recherches sur la didacticité des discours ordinaires*, *4*, 95–110.

Py, B. (2000). Didactique des langues étrangères et recherche sur l'acquisition. Les conditions d'un dialogue. *Études de Linguistique Appliquée*, *120*, 395–404.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Sousa, A., Ballier, N., Gaillat, T., Stearns, B., Zarrouk, M., Simpkin, A., & Bouyé, M. (2020). *From linguistic research projects to language technology platforms: A case study*. International Workshop on Language Technology Platforms IWLTP 2020 – co-located with LREC 2020, Marseille.

Tack, A., François, T., Roekhaut, S., & Fairon, C. (2017). Human and Automated CEFR-based Grading of Short Answers. *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 169–179. https://doi.org/10.18653/v1/W17-5018

Tibshirani, R. (1994). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288.

Tono, Y. (2013). Automatic extraction of L2 criterial lexicogrammatical features across pseudo-longitudinal learner corpora: using edit distance and variability-based neighbour clustering. In Bardel, C., Lindqvist, C. & Laufer, B. (eds.), *L2 Vocabulary Acquisition, Knowledge and Use: New Perspectives on Assessment and Corpus Analysis*. Online: The European Second Language Association, 149–176.

Vajjala, S. (2017). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*. *28, 79-105*. https://doi.org/10.1007/s40593-017-0142-3

van Rooy, B., & Schafer, L. (2003). An evaluation of three POS taggers for the tagging of the Tswana Learner English Corpus. In Archer, D., Rayson, P., Wilson, A. & McEnery, T. (eds.), *Proceedings of the Corpus Linguistics 2003 conference, 28-31 March 2003*. Lancaster: Lancaster University, 835–844.

Venant, R., & D'Aquin, M. (2019). Towards the Prediction of Semantic Complexity Based on Concept Graphs. In Lynch, C. F., Merceron, A., Desmarais, M. & Nkambou, R. (eds.), *12th International Conference on Educational Data Mining (EDM 2019)*, Canada, 188–197.

Volodina, E., Pilán, I., & Alfter, D. (2016). Classification of Swedish learner essays by CEFR levels. In Papadima-Sophocleous, S., Bradley, L. & Thouësny, S. (eds.), *CALL communities and culture – short papers from EUROCALL 2016*. Research-publishing.net, 456–461.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*. Second Language Teaching &

Curriculum Center, University of Hawaii at Manoa.

Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011). A New Dataset and Method for Automatically Grading ESOL Texts. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, 180–189.