TCFLE-8: a Corpus of Learner Written Productions for French as a Foreign Language and its Application to Automated Essay Scoring

Rodrigo Wilkens∗, Alice Pintard∗, David Alfter†, Vincent Foly‡, Thomas François∗

∗Cental, IL&C, UCLouvain, †University of Gothenburg, ‡France Éducation international

Abstract

Automated Essay Scoring (AES) aims to automatically assess the quality of essays. Automation enables large-scale assessment, improvements in consistency, reliability, and standardization. Those characteristics are of particular relevance in the context of language certification exams. However, a major bottleneck in the development of AES systems is the availability of corpora, which, unfortunately, are scarce, especially for languages other than English. In this paper, we aim to foster the development of AES for French by providing the TCFLE-8 corpus, a corpus of 6.5k essays collected in the context of the Test de Connaissance du Français (TCF - French Knowledge Test) certification exam. We report the strict quality procedure that led to the scoring of each essay by at least two raters according to the levels of the Common European Framework of Reference for Languages (CEFR) and to the creation of a balanced corpus. In addition, we describe how linguistic properties of the essays relate to the learners’ proficiency in TCFLE-8. We also advance the state-of-the-art performance for the AES task in French by experimenting with two strong baselines (i.e., RoBERTa and feature-based). Finally, we discuss the challenges of AES using TCFLE-8.

1 Introduction

Automated Essay Scoring (AES) aims to develop algorithms that can assess the quality of essays similarly to humans. The field may be traced back to the seminal work of Page (1966). Since then, several publications have been studying AES. In the late 1990’s, several functional AES systems were already available, either relying on Latent Semantic Analysis (e.g., Landauer et al. (1997)), NLP-extracted features combined with multiple regression (e.g., Burstein et al. (1998)) or Bayesian text classification (e.g., Rudner and Liang (2002)). As noted by Dikli (2006), a small amount of essays (less than 1000) could be enough for training such systems in some contexts. However, even collecting such a small corpus was difficult, as the essays need to be manually rated, and essays reliable assessment is a notoriously difficult task for humans (Wolfe et al., 2016).

Recent advances in AES have been made possible by Deep Learning (DL) approaches and large language models (Ramesh and Sanampudi, 2022). Prominent studies used embeddings (Alikaniotis et al., 2016), recurrent neural network (Taghipour and Ng, 2016), attention (Dong et al., 2017), and BERT-based architectures (Mayfield and Black, 2020). These approaches have also led to a growing need for large corpora.

Consequently, AES teams have turned their attention to learner corpus research, a branch of corpus linguistics providing large-scale, computerized, naturalistic learner production. Pioneering works such as the International Corpus of Learner English (ICLE) (Granger, 1993) and the European Science Foundation L2 Database (Perdue, 1993) demonstrated the potential of such learner data collections for Second Language Acquisition (SLA) research, but it is only recently that more learner corpora fitted for AES, i.e., large enough and annotated with different proficiency levels, were developed for various languages (Yannakoudakis et al., 2011; Blanchard et al., 2013; Geertz et al., 2014; Wisniewski et al., 2013; Mendes et al., 2016; Sakoda and Hosoi, 2018).

Unfortunately, there is no such large corpus for French, making the situation for French AES far from encouraging. The first systems thus relied on unsupervised approaches: Lemaire and Dessus (2001) used Latent Semantic Analysis to compare
native language (L1) of student essays with textbook passages, whereas AUTO-EVAL (Zaghouani, 2002) automatically captured several L1 essay features, which are heuristically combined. More recently, Parslow (2015a) trained a Naive Bayes classifier on a very small corpus of 200 essays written in foreign language (FL). Finally, Ranković et al. (2020) were the first to fine-tune BERT for FL French AES on more data, but they did not release it and only a single L1 is represented.

Therefore, in order to support the development of AES solutions for French, the need for a large and reliable corpus of written French essays becomes apparent. In this paper, we make two main contributions. First, we provide the community with the TCFLE-8 corpus\(^3\), composed of 6,569 learner essays, with 8 different languages of habitual use, scores, from at least 2 raters, for the 6 levels of the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001), and automatically annotated with +5k features. These essays were collected in the context of the official French Knowledge Test (TCF) exam, one of the main certification exams for French. Second, we provide solid baselines for future research in AES using TCFLE-8. This is the largest AES modeling study that has been done for French.

This paper is organized as follows. Section 2 presents the main characteristics and uniqueness of corpora used for AES. We then present the corpus compilation process, presenting the official certification exam which our corpus is based on and the essay selection process (Section 3). Next, in Section 4, we present the TCFLE-8 corpus, discussing its size, metadata and annotation. An exploration of TCFLE-8 for AES systems is presented in Section 5. Finally, final remarks are presented in Section 6. TCFLE-8 is freely available for research purposes.

2 Related Work

Developing a French corpus for AES means taking part in the field of learner corpus research, which, since its emergence in the late 1980s, gave rise to more than 200 learner corpora around the world\(^4\). Reviews on learner corpora (Gilquin, 2015; Granger et al., 2013) point out the prevalence of English as target language for more than half of the corpora, the rest focusing on German (Siemen et al., 2006; Gut, 2012; Belz, 2004), Spanish (Lozano, 2009; Cestero Mancera et al., 2002), French (Granger, 2003; Granfeldt et al., 2006), Italian (Di Nuovo et al., 2022) and others (Atwell and Alfaïfi, 2014; Wang et al., 2015; Martin et al., 2012). In terms of usual or native language of the learners (L1), the majority of learner corpora are mono-L1. Multi-L1 corpora are however favored today because they allow to study the influence of various L1s on the target language and they offer a wider degree of generalization.

Among this large body of written learner corpora, we will focus on two types relevant for this work: corpora used for AES and learner corpora targeting French as a foreign language.

2.1 Corpora for AES

Corpora built for AES can focus of specific dimensions, such as the organizational skill of essay writing (e.g., ICLE (Granger, 1993)) or the persuasive nature of the essay (e.g., Argument Annotated Essays (Stab and Gurevych, 2014)), but they usually address the global level of a learner production on a proficiency scale. One of the most commonly used scales for foreign languages is the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001), describing six levels of proficiency from A1 (beginner) to C2 (advanced). As the mapping between each essay and its proficiency level is critical in AES, it is best to use essays written in the context of official L2 certifications, as they benefit from strict rating procedures, usually with at least two professional raters grading each production. We distinguish these corpora, which we call candidate corpora, from other learner corpora containing productions collected in language classes or on web forums.

2.1.1 Candidate corpora

In addition to having more reliable proficiency ratings, candidate corpora also contain more varied learner profiles in terms of L1, age and background. The largest candidate corpus is the Cambridge Learner Corpus (CLC) (Nicholls, 2003) with more than 50 million words from 200,000 written productions and 138 different L1s. It was compiled from English exams of Cambridge Assessment English and two subparts of this corpus are available for research. First, the OpenCLC (Lexical Com-
puting Limited, 2017) is composed of more than 10,000 texts from candidates of 7 different L1s. The second available subpart of the CLC, First Certificate of English (CLC-FCE), contains 1,238 texts aligned with the CEFR (Yannakoudakis et al., 2011; Vajjala and Rama, 2018). Another corpus targeting English was released by Educational Testing Service, the ETS corpus of non-native written English or TOEFL11 (Blanchard et al., 2013). Initially compiled to perform L1 detection tasks, this corpus was later used in AES to explore both traditional machine learning (Rupp et al., 2019) and deep learning models (Nadeem et al., 2019). It contains 12,100 English essays written by TOEFL candidates of 11 non-English native languages. The essays are presented with their prompt and proficiency level given by ETS (low-medium-high).

Collaborations between certified testing organisations and research groups also developed for other European languages, resulting into three recent candidate corpora. The MERLIN corpus (Wisniewski et al., 2013) contains 2,290 written productions from standardized tests targeting German, Italian (TELc institute) and Czech (UIOP Institute). Its design allowed for cross-lingual AES experiments (Vajjala, 2018; Lehner et al., 2020; Bestgen, 2020; Caines and Buttery, 2020). The COPLE2 corpus (Mendes et al., 2016), containing 966 essays written in Portuguese (ICLP and CAPLE institutes), and the ASK corpus (Tenfjord et al., 2006b), with 1,936 texts written by candidates to the Norwegian Language Test, have also both been used for AES (del Rio et al., 2016; Berggren et al., 2019; Carlsen, 2012). Table 6 in Appendix A provides an additional detailed description of existing candidate corpora.

2.1.2 Corpora from language classes

Some corpora compiled from learner productions in language classes also prove to be suitable for AES tasks. For example, EFCAMDAT (Education First-Cambridge Open Language Database) (Geertzen et al., 2014) has been used in AES to investigate features related to the CEFR scale (Arnold et al., 2018), to classify based on errors (Ballier and Gaillat, 2016) or neural AES models (Kerz et al., 2021). To the best of our knowledge, this is the largest L2 corpus used in AES that does not come from certification exams. EFCAMDAT contains 83 million words from more than 1 million essays written by learners of Education First’s online English school. These essays span 16 levels traceable to the CEFR scale, and the prompts are level-specific (Geertzen et al., 2013).

AES experiments were also conducted for Spanish on CEDEL2, a learner corpus of more than 1 million words from 4,399 learners of 11 different L1s (Lozano, 2009), for Swedish on the SweLL corpus containing approximately 600 texts (Volodina et al., 2016) and for Japanese on the I-JAS corpus of texts written by 1000 learners of 12 different native languages (Sakoda and Hosoi, 2018). These experiments involve traditional machine learning work with features (del Río et al., 2016; Pilán and Volodina, 2018; Lee and Hasebe, 2020) or deep learning (Lilja, 2018; Ruan, 2020; Hirao et al., 2020a).

2.2 Learner corpora targeting French

To the best of our knowledge, there are no candidate corpora for French. Most learner corpora targeting French were compiled to study interlanguage5 (Selinker, 1972). They were collected from language courses at university, so the levels represented are mainly intermediate and advanced. The French Interlanguage Database (Granger, 2003) contains 450,000 words. Other corpora designed for interlanguage investigation include the Learner Corpus French (Vanderbauwhede, 2012), containing 500,000 words, and the Chy-FLE/HellasFLE (Valetopoulos and Zajac, 2012), containing 150,000 words. The Corpus Interlangue (Gaillat and Roa, 2020), a written/spoken and bilingual corpus, contains texts and interviews from 115 students. The Corpus Écrit de Français Langue Étrangère (Granfeldt et al., 2006) approaches learners interlanguage in the language development sequences. It is the only corpus representative of all proficiency levels for French, and it contains 100,000 words. It has been used for AES to find the features most correlated with CEFR levels (Parslow, 2015a). Finally, the French part of the Word Reference Corpus (Berdicevsksis, 2020) constitutes the largest learner corpus for French with 4 million words from forum posts on the Word Reference website. It has been used to study contact-induced simplification, but despite its considerable size, it was not used for AES, because it is noisy and text levels have not been evaluated. More information on learner corpora targeting French is presented in Table 6 in Appendix A.

---

5Interlanguage describes the unique linguistic organisation developed by a foreign language learner, which presents some features of previously acquired language and may overgeneralize L2 patterns.
3 Corpus compilation

3.1 Data collection

TCFLE-8 being a candidate corpus for French, it has been collected by one of the agencies carrying out official certification in L2 French: France Education International (FEI). FEI is a French agency under the supervision of the Ministry of National Education and Youth. With a workforce of over 250 employees and a network of more than 1,000 experts, FEI acts in various fields of cooperation in education and training and contributes to the promotion of the French language and the French-speaking world. FEI offers a wide range of certifications in French aligned with the six CEFR levels: initial diploma in French language (DILF), diploma in French language studies (DELF), diploma in advanced French language studies (DALF) and French knowledge test (TCF). Around 650,000 candidates take one of these examinations on an annual base in more than 180 countries.

As its name implies, TCFLE-8 is based on the TCF, a linear test aligned with the six CEFR levels. The TCF is used mainly for academic studies, migration purposes and citizenship. Its written component, made up of three independent tasks, is taken annually by 120,000 candidates, 60% of which sit their exam on computer.

The correction is performed by professional raters. FEI has a pool of about 100 raters, recruited on occupational profiles (experienced teachers, previous experience for rating with French). Applicant raters take a psychometrically-calibrated rating competence test for writing and attend a two-day training. At the end of this procedure, the recruitment is confirmed or not. To ensure reliability in the long term, reliability indices of raters are assessed annually, and a decision is made regarding whether to retain them in the pool. In addition, to ensure reliability at the candidate level, FEI adopts a double rating approach. In case of discrepancy, a third rater is called to independently rate the 3 productions. The final level of the candidate is established based on the frequency of the CEFR levels given to the three candidate’s productions.

To identify the CEFR level, the raters use adapted CEFR descriptors and scales. The descriptors (and the rating) are holistic, although each descriptor is aiming at linguistic (organisational), pragmatic and sociolinguistic dimensions and their related criteria. Until now, language test providers used both analytical and holistic scales (Hamp-Lyons, 1995). There is no clear consensus on the superiority of one type of scale in terms of reliability and efficiency (Ono et al., 2019).

Language competence is multidimensional (Bachman, 1990; Bachman and Palmer, 2010; Oller and Hinofotis, 1980; Vollmer and Sang, 1983) and is a measurable skill (Vollmer and Caroll, 1983). Measuring writing skill implies considering various facets: candidate proficiency, rater leniency/harsness and difficulty of the task. To this aim, “Many-facet Rasch measurement (MFRM) is a psychometric approach that establishes a coherent framework for drawing reliable, valid, and fair inferences from rater-mediated assessments, thus answering the problem of fallible human ratings” (Eckes, 2009). Therefore, we applied MFRM to the FEI dataset of TCF exams in order to identify and avoid the fallible human ratings in the data set.

3.2 Data cleaning

The original data collected by FEI had to be cleaned at various levels. First, outlier identification consisted in removing candidates’ responses that did not achieve the A1 level, were copies of the prompt, too short/long, or off-topic. Next, we leverage the Rasch information in the dataset to detect texts for which human raters might have failed to provide a reliable judgment. We compared FEI raters’ original scores and the scores adjusted by the Rasch method, using standardized residuals. After an empirical evaluation, we removed all essays with a standardized residual value greater than 4. In addition, we also dropped essays with a low confidence assessment (e.g., candidates that are on the borderline between levels). To accomplish this, we removed all cases where both raters disagree with each other and with the candidate’s final score, and we also removed the cases where there is a distance of three CEFR levels between the lowest and the highest ratings. After this process, we set the essay score as the candidate’s CEFR level when at least one of the raters assigned that level to the essay. Alternatively, if both raters agreed with the essay’s score, we duly assigned this level to the essay. Any essay that does not fit any of these two criteria has been removed.

It has to be mentioned that the rating of the set of the 3 tasks is done by the same rater, thus not being independent.  

In our empirical evaluation, we explored four standardized residue values (2, 3 and 4), observing that around 5.6% of the corpus has a standardized residue of 2, 0.9% has a value of 3 and 0.4% a value of 4.
After outlier removal, the next step was to get a representative sample from the set of TCF essays available. For a fair representation, the level of the text is an obvious variable to control. In addition, we controlled for the language of habitual use, aiming for a representation of the most frequent languages. As the top five were all European ones and the 6th was Kabyle, an Afro-Asiatic language, we also included Chinese and Japanese to get a better representation of various typological families of languages. Thus, we launched a random sampling controlling for the 6 CEFR levels and the candidate’s language of habitual use. To apply this algorithm, we set up an objective function that approximates the CEFR scores distribution by language. In order to reflect the distribution in the whole dataset, we divided the current language into frequency bands: very frequent (English and Arabic), frequent (Spanish, Kabyle, Portuguese and Russian) and infrequent (Chinese and Japanese) languages. For a description of the resulting corpus see Section 4.

3.3 (Pseudo-)anonymization

Candidates sometimes include personal information in their essays. While this does not pose a problem for the assessment, it can expose candidates when the texts become public. This exposure is generally tackled with anonymization methods (e.g., Wisniewski et al. (2013); Mendes et al. (2016); Tenfjord et al. (2006a); Gablasova et al. (2019); Rakhilina et al. (2016)) or pseudo-anonymization (e.g., (Glaznieks et al., 2020; Preradovic et al., 2015; Rosen et al., 2020; Dirdal et al., 2022)) in the literature on learner corpora. Typically, these processes capture names (e.g., Gablasova et al. (2019); Preradovic et al. (2015); Rosen et al. (2020); Rakhilina et al. (2016)), but sometimes they also capture other information, such as location and date (e.g., Glaznieks et al. (2020); Tenfjord et al. (2006a); Wisniewski et al. (2013) ), geo-data (e.g., Volodina et al. (2019)) and language-specific substitutions (e.g., Wisniewski et al. (2013)). In our work, we decided to provide anonymous and pseudo-anonymous versions of TCFLE-8. The latter is intended to provide a more natural text, but pseudo-anonymization may introduce grammatical errors (e.g., wrong contractions).

This work uses the MAPA tool (Gianola et al., 2020) for (pseudo-)anonymization. With this tool, we target 7 entities: names, address (i.e., country, city, building, territory and place), date (i.e., day of week, month, year and day), e-mail, organization, amount and phone. After the pseudo-anonymization, we assessed its quality. During this process, we noticed some consistent flaws in the tool that were corrected in the corpus to improve its quality. The observed issues consisted of an overanonymization of words at sentence beginnings when predicated by a name, and an omission to replace email addresses.

4 The TCFLE-8 corpus

At the end of the compilation process, the final TCFLE-8 corpus comprises 6,569 essays (581,333 words). Some figures about the corpus size by CEFR levels are shown in Tables 1, 2, 3 and 4. It is expected that beginner-level essays tend to be shorter than the other ones (Fraser et al., 1998). Moreover, the extreme levels (A1 and C2) are less represented in the corpus. This might be caused by two factors: (1) few A1-level learners seek an official language exam since this level is rarely sufficient for official purposes (e.g., employment and visa requirements), and (2) reaching the C2 level in a foreign language is extremely difficult.

<table>
<thead>
<tr>
<th>Level</th>
<th>#essays</th>
<th>%essays</th>
<th>avg #wrd (stdev)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>689</td>
<td>10.49</td>
<td>69.78 (34.02)</td>
</tr>
<tr>
<td>A2</td>
<td>1375</td>
<td>20.93</td>
<td>91.69 (44.80)</td>
</tr>
<tr>
<td>B1</td>
<td>1466</td>
<td>22.52</td>
<td>119.11 (49.89)</td>
</tr>
<tr>
<td>B2</td>
<td>1427</td>
<td>21.72</td>
<td>133.61 (44.92)</td>
</tr>
<tr>
<td>C1</td>
<td>1127</td>
<td>17.16</td>
<td>133.92 (45.91)</td>
</tr>
<tr>
<td>C2</td>
<td>485</td>
<td>7.38</td>
<td>138.92 (48.45)</td>
</tr>
<tr>
<td>Total</td>
<td>6569</td>
<td>100.0</td>
<td>119.67 (50.31)</td>
</tr>
</tbody>
</table>

Table 1: Description of the TCFLE-8 corpus by CEFR level: number of essays, percentage, and mean and standard deviation of word number per essay.

Table 2 indicates the number of essays distinguishing the gender. It is interesting to note that about 58% of the sample is composed of women and this proportion is not the same at all level. Table 3 shows the essays by three tasks in the TCF exam, where there is no general difference between one task and the others. Finally, Table 4 picture the amount of essays in the different languages of habitual use.

---

8The language of habitual use is the language the candidate indicates as the one they usually use.
9https://gitlab.com/MAPA-EU-Project/
10The evaluation scores are presented in Section B.
Comparing TCFLE-8 to existing corpora (see Table 6 in Appendix A), it is the largest French learner corpus suitable for AES – both in size and L1 representation –, the third largest candidate corpus to our knowledge and its annotation layers provide the richest information (see Section 4.2). It also covers all 6 CEFR levels.

### 4.1 Metadata

As a complement to the text of the essays and their CEFR scores, assigned according to the procedure described in Section 3.2, the TCFLE-8 corpus provides information about the candidate who wrote the essay and the essay prompt.

Regarding the candidates’ information, their gender and language of habitual use are provided. The candidates communicate this information when they register for the TCF exam. Overall, the corpus contains slightly more women than men (58% vs 41%), see Table 2. As for the language of habitual use, the corpus covers 8 languages, as described in Section 3.2: English, Arabic, Spanish, Russian, Portuguese, Kabyle, Chinese, and Japanese, respectively, with 917, 906, 904, 889, 872, 866, 681, and 534 essays (see Table 4).

The CEFR level achieved by each candidate considering the three written productions is also reported. This is the official CEFR level assigned to the candidate for the written part of the TCF exam.

### 4.2 Essay annotation

In addition to the above metadata, TCFLE-8 also includes a linguistic annotation layer aimed to describe the learners’ proficiency. This annotation was automatically performed using the FABRA toolkit (Wilkens et al., 2022). It allows computing the distribution of over 400 linguistic variables grouped by family of related variables (e.g., lexical diversity, and lexical frequency). These distributions are aggregated using 18 statistical descriptors, which results in more than 5k annotations per essay. In addition, we extended the existing FABRA features by including others related to SLA. In particular,

---

<table>
<thead>
<tr>
<th>(CEFR) Level</th>
<th>Men</th>
<th>Women</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>394</td>
<td>295</td>
</tr>
<tr>
<td>A2</td>
<td>574</td>
<td>801</td>
</tr>
<tr>
<td>B1</td>
<td>596</td>
<td>870</td>
</tr>
<tr>
<td>B2</td>
<td>537</td>
<td>890</td>
</tr>
<tr>
<td>C1</td>
<td>441</td>
<td>686</td>
</tr>
<tr>
<td>C2</td>
<td>198</td>
<td>287</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>2740</td>
<td>3829</td>
</tr>
</tbody>
</table>

Table 2: Number of essays according to the gender, per (CEFR) level

<table>
<thead>
<tr>
<th>(CEFR) Level</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>225</td>
<td>223</td>
<td>241</td>
</tr>
<tr>
<td>A2</td>
<td>426</td>
<td>413</td>
<td>536</td>
</tr>
<tr>
<td>B1</td>
<td>447</td>
<td>475</td>
<td>544</td>
</tr>
<tr>
<td>B2</td>
<td>485</td>
<td>485</td>
<td>457</td>
</tr>
<tr>
<td>C1</td>
<td>358</td>
<td>431</td>
<td>338</td>
</tr>
<tr>
<td>C2</td>
<td>156</td>
<td>173</td>
<td>156</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>2097</td>
<td>2200</td>
<td>2272</td>
</tr>
</tbody>
</table>

Table 3: Number of essays according to the task number

<table>
<thead>
<tr>
<th>Lang.</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>JPN</td>
<td>8</td>
<td>135</td>
<td>171</td>
<td>170</td>
<td>48</td>
<td>2</td>
</tr>
<tr>
<td>CHI</td>
<td>34</td>
<td>165</td>
<td>244</td>
<td>189</td>
<td>45</td>
<td>4</td>
</tr>
<tr>
<td>SPA</td>
<td>124</td>
<td>187</td>
<td>175</td>
<td>182</td>
<td>178</td>
<td>58</td>
</tr>
<tr>
<td>ARA</td>
<td>135</td>
<td>160</td>
<td>163</td>
<td>153</td>
<td>160</td>
<td>135</td>
</tr>
<tr>
<td>POR</td>
<td>102</td>
<td>187</td>
<td>182</td>
<td>191</td>
<td>172</td>
<td>38</td>
</tr>
<tr>
<td>ENG</td>
<td>125</td>
<td>163</td>
<td>167</td>
<td>165</td>
<td>169</td>
<td>128</td>
</tr>
<tr>
<td>RUS</td>
<td>103</td>
<td>198</td>
<td>183</td>
<td>196</td>
<td>180</td>
<td>29</td>
</tr>
<tr>
<td>KAB</td>
<td>58</td>
<td>180</td>
<td>181</td>
<td>181</td>
<td>175</td>
<td>91</td>
</tr>
</tbody>
</table>

Table 4: Number of essays according to the language of habitual use (language code follows ISO639-2)

The Quadratic Weighted Kappa (QWK) between the candidate’s level and the CEFR-level of the essay is 0.98. It is expected that this value should be high, but not equal to 1, due to the cases where candidates cannot maintain a consistent level of essay quality. In addition, the scores assigned by the FEI raters in the double rating procedure are also available. They have a QWK of 0.71 (correlation of 0.93) with each other and 0.84 (correlation of 0.85) with the CEFR-level of the essay.

Finally, the prompt and its position in the sequence of three TCF tasks are also reported. This information, which relates to the exam, contextualizes the essay’s input and the grading sequence (described in Section 3.1). Regarding the prompt position, it is balanced in TFCFL-8 (32% essays for the first task, 33% for the second, and 35% for the third). Table 3 shows the distribution of the three tasks across the six CEFR levels.\(^{11}\)

---

\(^{11}\)We calculated the correlation – Spearman for continuous variables or Point-biserial for binary variables – between the CEFR score and the metadata described above to identify possible biases in the scores; all correlations were between 0.038 and 0.09. This analysis confirms that the sampling process did not induce unexpected biases.
In order to better characterize how linguistic properties present in TCFLE-8 are associated with the learners’ proficiency, we computed Spearman’s correlations between each of above feature (gathered by the families in Wilkens et al. (2022)) and the essays’ CEFR level. In the process, we dropped features with correlations lower than 0.4 or p-values higher than 0.05. Next, as many features are variants of each other, we calculated the correlation matrix within each family to identify redundant features (i.e., an absolute correlation above 0.90). Finally, for each set of similar features, we considered only the one most correlated with the CEFR level. After this procedure, we kept 119 correlated features. In Figure 1, we show their distribution by family of variable.\(^{13}\)

Our analysis of the selected features highlighted linguistic properties of essays already reported in studies investigating foreign language writing. For example, measures of word length have been known to be good predictors of proficiency level for English (Ferris, 1994; Grant and Ginther, 2000), for Swedish (Pilán and Volodina, 2018), for Japanese (Hirao et al., 2020b) and for French Parslow (2015b). As regards lexicon, diversity measures (e.g. type/token ration) correlate with proficiency in English (Lu, 2012; Vajjala, 2018), whereas sophistication measures based on word frequencies have yielded similar results in a number of studies: more proficient writers use, on average, fewer frequent words (Laufer and Nation, 1995; Attali and Burstine, 2006; Crossley and McNamara, 2012; Guo et al., 2013). Finally, the most discriminating feature in TCFLE-8, namely the error-rate, is also one of the most correlated features to proficiency levels in the CLC-FCE and TOEFL11 (Yannakoudakis et al., 2011; Vajjala, 2018). In addition, while this analysis confirms existing research findings in AES, it also points out that TCFLE-8 may be helpful for new SLA studies. Indeed, we also provided several features explored in other acquisition-related fields, such as the OLD20 (measuring orthographic similarity) (Coltheart et al., 1977; Yarkoni et al., 2008), which are significant in our corpus but had not been linked to L2 writing proficiency so far, to the best of our knowledge.

5 AES for French

In this section, we analyze the applicability of the TCFLE-8 corpus for training AES systems. For this purpose, we explore two approaches: deep learning, since most AES systems relied on neural networks (Ramesh and Sanampudi, 2022), and feature-based machine learning.

We split the anonymized corpus with 80% for training, 10% for validation, and 10% for testing, stratifying by score and language. In addition, to explore the impact of model initialization, we performed 5 repetitions of the training process; in each one, we adjusted the test set so that it does not overlap with the others. We performed a hyperparameter exploration using the accuracy on the validation set.

For the deep learning model, we used CamemBERT (Martin et al., 2020), a RoBERTa-based model for French. As for the hyperparameters\(^{14}\), we use a learning rate of 5e-5 and an early stop of 5. For machine learning, we use the XGBoost\(^{15}\) and

\(^{12}\)https://pypi.org/project/language-tool-python/

\(^{13}\)Table 9 in the appendices presents the list of all correlated features and their correlation values.

\(^{14}\)The hyperparameters search explored 1e-4, 5e-5, 1e-5, 5e-6 and 1e-6 as learning rate, 1, 3, 5, 7 and 10 as early stop patience, searching up to 40 epochs.

\(^{15}\)The hyperparameters used for XGBoost and the values explored are gbtree as booster, alternatively exploring gbtree, gbrf, gblinear and dart, 0.3 as subsample, from 0.3 and 0.6. 3 as
logistic regression\textsuperscript{16} as a feature-based baseline. These were trained using the 119 features extracted using the method described in Section 4.2. The evaluation of these models is shown in Table 5.

In order to characterize human level performance on the task, we report standard AES evaluation metrics for the human raters (column “raters” in Table 5), namely accuracy, adjacent accuracy, F1-score, and QWK.\textsuperscript{17} Those metrics were calculated by a direct comparison between the ratings of one of the two evaluators and the reference CEFR levels for each essay. Those results show that the task of identifying the CEFR level of an essay is hard, even for humans. However, the adjacent accuracy of 0.99 clearly shows that the identification gap is typically up to one level. In the same direction, the QWK points out that once the ordinality existing between CEFR levels is considered, the agreement among raters is remarkably strong. As expected, none of our models achieved results competitive with human performance.\textsuperscript{18}

Among the AES models explored, the transformer-based CamemBERT achieved the best values. Despite this performance, it can be seen that there is still room for improvement when comparing the results with the evaluation by experts (column raters). Considering that the transformers model performs in a range between raters and XGBoost, it is interesting to remark that the transformers model is closer to the raters’ performance when we consider the ordinality relation between levels (i.e., QWK and Accuracy\textsubscript{Adjacent}). Focusing on the ability of models to discriminate specific levels, the fine-tuned version of CamemBERT emerged as a model of better performance. Moreover, the logistic model is clearly a weak baseline. Interestingly, at the C2 level, which was the most challenging for all three models, XGBoost suffers from a catastrophic failure, achieving even lower performance than the Logistic model. This general weak performance is not entirely surprising, as texts at this level tend to explore language idiosyncrasies, to be precise, to have a very coherent and organized structure, etc. In contrast, the beginner levels (i.e., A1 and A2), for which transformers and XGBoost models had close results, is characterized by texts with simple vocabulary and grammatical structures.\textsuperscript{19}

As TCFLE-8 is a new corpus for the French language, we cannot fairly compare our results with previous works, due to the considerable difference in corpus size. In the French AES literature, we identified only two papers focusing on L2 proficiency identification. First, Parslow (2015a), who used a corpus of 200 essays to train a Naive Bayes classifier and reported F1-scores ranging from 0.51 to 0.74 for the levels A1 to B2. Second, Ranković et al. (2020) used CamemBERT intermediate layers as features to predict level in a corpus of 100 essays and reported MSE ranging from 0.35 to 0.55.

6 Final Remarks

In this work, we presented TCFLE-8, a corpus of 6,569 candidates’ essays written during the French knowledge test (TCF), with 8 different languages of habitual use. This paper described the data gathering by France Education International (FEI), data cleaning, anonymization, and annotation performed to compile this corpus, which is the largest French corpus targeting French as a foreign language for AES. This corpus, along with its metadata (i.e., essays, metadata and annotation) is available to the community. We also described the learners’ proficiency in the corpus using numerous linguistic variables related to SLA. This description confirms that these linguistic features could capture developmental patterns in TCFLE-8 in a similar fashion to other learner corpora.

Exploring TCFLE-8 for AES, we applied different machine learning algorithms. CamemBERT appears to be more accurate and XGBoost, a feature-based model, achieved similar results at beginner level. This raises a question about what features should explored for better describing the intermediate and advanced levels. Interestingly, part of this answer may come from the transformer model

---

\textsuperscript{16} For the logistic regression, we explored the following hyperparameters: penalty from l2 or none, C from 0, 1, 10, max interaction from 100 or 300, and multi class process from multinomial or one-vs-rest. After this short exploration, we set the solver as lbfgs, l2 as the penalty, the C and the max interaction as 100 and 300, and the class processing as one-vs-rest.

\textsuperscript{17} For a transparent presentation of our results, Appendix C shows the confusion matrices of the transformer-based model.

\textsuperscript{18} Note that the scores in the “raters” column are inflated, because the rating assigned by each rater contributes to the final CEFR score of each essay.

\textsuperscript{19} We compared the results of training the models on the anonymized corpus with the corresponding models trained on the original corpus (before anonymization) and no statistical difference was identified.
Table 5: Average and standard deviation of the evaluation score using the TFCFL-8 for AES and the performance of human raters

<table>
<thead>
<tr>
<th>Metric</th>
<th>CamemBERT</th>
<th>XGBoost</th>
<th>Logistic</th>
<th>Raters</th>
</tr>
</thead>
<tbody>
<tr>
<td>QWK</td>
<td>0.88 (0.01)</td>
<td>0.79 (0.02)</td>
<td>0.69 (0.02)</td>
<td>0.93 (0.01)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.57 (0.01)</td>
<td>0.46 (0.01)</td>
<td>0.37 (0.01)</td>
<td>0.76 (0.01)</td>
</tr>
<tr>
<td>Accuracy/Adjacent</td>
<td>0.98 (0.01)</td>
<td>0.92 (0.02)</td>
<td>0.80 (0.01)</td>
<td>0.99 (0.01)</td>
</tr>
<tr>
<td>F1_weighted</td>
<td>0.56 (0.01)</td>
<td>0.46 (0.02)</td>
<td>0.36 (0.02)</td>
<td>0.76 (0.01)</td>
</tr>
<tr>
<td>A1_F1</td>
<td>0.63 (0.01)</td>
<td>0.59 (0.04)</td>
<td>0.54 (0.06)</td>
<td>0.76 (0.02)</td>
</tr>
<tr>
<td>A2_F1</td>
<td>0.57 (0.04)</td>
<td>0.53 (0.01)</td>
<td>0.40 (0.05)</td>
<td>0.76 (0.03)</td>
</tr>
<tr>
<td>B1_F1</td>
<td>0.56 (0.04)</td>
<td>0.45 (0.05)</td>
<td>0.32 (0.02)</td>
<td>0.75 (0.01)</td>
</tr>
<tr>
<td>B2_F1</td>
<td>0.56 (0.04)</td>
<td>0.43 (0.03)</td>
<td>0.34 (0.03)</td>
<td>0.76 (0.02)</td>
</tr>
<tr>
<td>C1_F1</td>
<td>0.56 (0.04)</td>
<td>0.42 (0.03)</td>
<td>0.30 (0.05)</td>
<td>0.77 (0.02)</td>
</tr>
<tr>
<td>C2_F1</td>
<td>0.48 (0.09)</td>
<td>0.19 (0.07)</td>
<td>0.31 (0.02)</td>
<td>0.80 (0.04)</td>
</tr>
</tbody>
</table>

Itself (e.g. through a probing approach (Tenney et al., 2019)).

Finally, TCFLE-8 was portrayed in this paper as a corpus for French AES but its properties allow for different applications in NLP, SLA and educational studies. It may be a valuable corpus for pedagogical material development, whether they be dictionaries (Longman, 2002), activities focusing on common learner difficulties and errors (Kaszubski, 1998; Reppen, 2010), computer-assisted language learning software (Granger, 2003) or L2 writing aids (Link et al., 2014). Activities of data-driven learning in language class (Friginal, 2018) could also take advantage of this corpus. With 8 different languages of habitual use, this corpus could also be beneficial for cross-linguistic studies such as transfer mechanisms and L1 influence on L2 production (Golden et al., 2017; Werner et al., 2020), and for automatic native language identification (Tetreault et al., 2013). Another possible application for this new corpus is the one of errors detection and correction (Dahlmeier et al., 2013), that we are currently investigating as future work on TCFLE-8.

7 Limitations

Several normalization steps were applied in order to develop a coherent corpus for AES, aiming to compile a high quality corpus illustrating the proficiency levels with learner productions on which professional raters would agree. As a consequence, some potentially interesting cases were removed. This concerns for example texts on the borderline between two levels. Although they are interesting cases as they could support studies on understanding of the level boundaries, we opted for a corpus that represents the texts of each level. TCFLE-8 is a corpus designed for supporting the research in French as a foreign language, including AES. In this work, the focus is on the corpus compilation. Despite the initial tests performed here, our goal does not include an exhaustive verification of the corpus’ applications nor an evaluation of various AES approaches. Finally, we do not intend nor recommend using TCFLE-8 for a fully-automated evaluation environment but to improve writing assessment in French as a foreign language.

Acknowledgements

This research has been funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under the grant MIS/PGY F.4518.21 and by a research convention with France Éducation International. Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region

References


Taylor Arnold, Nicolas Ballier, Thomas Gaillat, and Paula Lissón. 2018. Predicting CEFRL levels in...


Aivars Glaznıyeks, Jennifer-Carmen Frey, Maria Stopnıy, Lorenzo Zanasi, and Lionel Nicolas. 2020. LEONIDE-Longitudinal Learner Corpus in Italiano, Deutsch and English 1.0.


Sylviane Granger, Gaëtanelle Gilquin, and Fanny Meunier. 2013. Twenty Years of Learner Corpus Research: Looking back, Moving ahead.


John W. Oller, Jr and F. B. Hinofotis. 1980. Two measures of writing ability : factor analytic studies of a variety of language subtests.

Masumi Ono, Hiroyuki Yamanishi, and Yuko Hijikata. 2019. Holistic and Analytic Assessments of the


Nives Preradovic, Monika Berač, and Damir Boras. 2015. *Learner Corpus of Croatian as a Second and Foreign Language*.


Katrin Wisniewski, Karin Schöne, Lionel Nicolas, Chiara Vettori, Adrian Boyd, Detmar Meurers, Andrea Abel, and Jirka Hana. 2013. MERLIN: An online trilingual learner corpus empirically grounding the European Reference Levels in authentic learner data.


A Description of the learner and candidate corpora

Table 6 in this appendix provides a summary of candidate corpora collected from L2 certification exams and learner corpora targeting French for a comparison with TCFLE-8.

<table>
<thead>
<tr>
<th>CANDIDATE CORPUS FROM L2 CERTIFICATIONS</th>
<th>L2</th>
<th>NUMBER OF L1</th>
<th>LEVEL</th>
<th>NUMBER OF TEXTS</th>
<th>TESTING INSTITUTION</th>
<th>ANNOTATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Cambridge Learner Corpus (Open CLC)</td>
<td>ENG</td>
<td>7</td>
<td>all levels</td>
<td>10,000</td>
<td>Cambridge Assessment English</td>
<td>auto: POS</td>
</tr>
<tr>
<td>CLC - First Certificate English (CLC-FCE)</td>
<td>ENG</td>
<td>-</td>
<td>all levels</td>
<td>1,238</td>
<td>Cambridge Assessment English</td>
<td>auto: POS and syntactic manual: errors</td>
</tr>
<tr>
<td>ETS Corpus of Non-Native Written English</td>
<td>ENG</td>
<td>11</td>
<td>all levels</td>
<td>12,100</td>
<td>Education Testing Services (TOEFL)</td>
<td>raw</td>
</tr>
<tr>
<td>ASK</td>
<td>NNO</td>
<td>10</td>
<td>B1-B2</td>
<td>1,936</td>
<td>Norwegian Language Test</td>
<td>auto: POS manual: errors, syntactic</td>
</tr>
<tr>
<td>COPLE2</td>
<td>POR</td>
<td>14</td>
<td>A1-C1</td>
<td>966</td>
<td>CAPLE ICLP</td>
<td>auto: POS manual: errors</td>
</tr>
<tr>
<td>TCFLE-8</td>
<td>FRE</td>
<td>8</td>
<td>all levels</td>
<td>6,500</td>
<td>France Education International</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>LEARNER CORPUS TARGETING FRENCH</th>
<th>L2</th>
<th>L1</th>
<th>LEVEL</th>
<th>COLLECTION CONTEXT</th>
<th>NUMBER OF WORDS</th>
<th>ANNOTATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>French Interlanguage Database (FRIDA)</td>
<td>FRE</td>
<td>ENG DUT others</td>
<td>int - adv</td>
<td>450,000</td>
<td>language class (university)</td>
<td>manual: errors</td>
</tr>
<tr>
<td>Learner Corpus French</td>
<td>FRE</td>
<td>DUT</td>
<td>B2-C1</td>
<td>500,000</td>
<td>language class (university)</td>
<td>-</td>
</tr>
<tr>
<td>Chy-FLE Hellas-FLE</td>
<td>FRE</td>
<td>GRE</td>
<td>int - adv</td>
<td>150,000</td>
<td>language class (university)</td>
<td>manual: grammatical constituents order</td>
</tr>
<tr>
<td>Corpus Interlangue (CIL)</td>
<td>FRE</td>
<td>ARA Madarin ENG SPA FRE</td>
<td>B1-C1</td>
<td>(115 txt)</td>
<td>texts, read aloud and interviews from 115 students (university)</td>
<td>no annotation</td>
</tr>
<tr>
<td>Corpus Ecrit de Français Langue Etrangère</td>
<td>FRE</td>
<td>SWE</td>
<td>deb - adv</td>
<td>100,000</td>
<td>language class (high-school)</td>
<td>auto: POS manual: errors</td>
</tr>
<tr>
<td>Word Reference Corpus</td>
<td>FRE</td>
<td>ARA SPA ITA</td>
<td>-</td>
<td>not evaluated</td>
<td>FFL: 4M. forum posts of Word reference website</td>
<td>no annotation</td>
</tr>
<tr>
<td>Dire Autrement</td>
<td>FRE</td>
<td>ENG</td>
<td>int - adv</td>
<td>50,000</td>
<td>language class (university)</td>
<td>manual: lexical errors</td>
</tr>
<tr>
<td>TCFLE-8</td>
<td>FRE</td>
<td>ENGLARA SARPUS, POR,KAB, CHLJPN</td>
<td>A1-C2</td>
<td>580,000</td>
<td>written production of TCF certification (France Education International)</td>
<td>auto: text-level annotation</td>
</tr>
</tbody>
</table>

Table 6: Candidate corpora and French learner corpora

B (Pseudo-)anonymization

For (pseudo-)anonymization, we start by applying the MAPA tool in the entire corpus. Next, we followed Volodina et al. (2019) by selecting 200 random texts and evaluating them manually to assess MAPA’s output. In addition, we controlled the same amount of text from each level because different levels can affect the system differently. However, contrarily to Volodina et al. (2019), we also evaluate whether anonymization is appropriate. The reason for this stricter approach was to measuring the of distorting caused by the (pseudo-)anonymization step.

The MAPA’s evaluation was carried out by two independent French native speaker. Each one evaluated 100 essays. Later, a third evaluator double-checked the 200 essays searching for inconsistencies, which were fixed after discussion with the other evaluators. During this assessment, we identify standard errors. These errors, described in Section 3.3, were automatically corrected after we identified their patterns of occurrence.
The results of this evaluation is shown in Table 7. The first observation is about the ability to fully identify an entity where MAPA presents difficulty. However, we point out that partial anonymization is already considered correct. In addition, a small number of errors are caused by the entity type, as exemplified by the close scores in the partial matching columns in the table, where the only distinction is the consideration of the entity and span or just the span. In this evaluation, we highlight two scores: accuracy and F2. The first takes into account the words that have been correctly identified as non-anonymized. The second, on the other hand, is a variation of the F-score where recall receives a greater weight. F2 represents the interest of coverage but without a significant loss in precision. Given the difference between these scores, and the recall and precision values, we notice that MAP tends to overdo, identifying more terms than needed for anonymization. Although the anonymization method generates an abundance of edits in the text, it ensures quality in the process and in the protection of the privacy of the writers.

<table>
<thead>
<tr>
<th></th>
<th>span-based</th>
<th>entity-based</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Partial match</td>
<td>Exact match</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td>0.97</td>
<td>0.96</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>0.55</td>
<td>0.47</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>0.86</td>
<td>0.68</td>
</tr>
<tr>
<td><strong>F1</strong></td>
<td>0.67</td>
<td>0.56</td>
</tr>
<tr>
<td><strong>F2</strong></td>
<td>0.77</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Table 7: Result of the anonymization evaluation

C Confusion matrices of transformer-based AES model

Table 8 shows the results for each of the 5 repetitions (see Section 5) of the AES model based on transformers (column CamemBert in Table 5).
<table>
<thead>
<tr>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>48</td>
<td>26</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>51</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>57</td>
<td>65</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>30</td>
<td>87</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>11</td>
<td>100</td>
<td>37</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>29</td>
<td>76</td>
<td>38</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>21</td>
<td>99</td>
<td>22</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td>63</td>
<td>55</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>44</td>
<td>53</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>31</td>
<td>19</td>
</tr>
</tbody>
</table>

Table 8: Predictions of the 5 repetitions of the CamemBert models fine-tuned to TCFLE-8

C.1 Correlated features

In this section, we list the features identified by the feature selection method presented in Section 4.2. Table 9 list as features and their correlations, while Figure 2 plots the distribution of some feature across the six CEFR levels.
Figure 2: Violin plot of some of the top correlated features through the 6 CEFR levels

(a) Average number of syllables per word
(b) Average ratio of affixes
(c) Squared type/token ration
(d) 1st quartile of Words in FLELex resource for A1 CEFR level
(e) Average of A1 level lemmas according to (Pintard and François, 2020)
(f) Average errors automatically identified by Language Tool
(g) Average of the mean orthographic Levenstein distance (based on Lexique3)
(h) Average number of surface form words in the top 1000 words of Lexique3
(i) Word probability, based on Lexique3
<table>
<thead>
<tr>
<th>Variable Family</th>
<th>List of features and their correlations with essay CEFR level</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Graded features</strong></td>
<td>LEXgrdFMLA1&lt;sub&gt;std&lt;/sub&gt; (0.463), LEXgrdFSOOUA1&lt;sub&gt;10P&lt;/sub&gt; (-0.41), LEXgrdFA2&lt;sub&gt;min&lt;/sub&gt; (-0.416), LEXgrdFA1&lt;sub&gt;10P&lt;/sub&gt; (-0.529), LEXgrdFSOOUC1&lt;sub&gt;var&lt;/sub&gt; (0.402), LEXgrdFA2&lt;sub&gt;20P&lt;/sub&gt; (-0.532), LEXgrdBA1&lt;sub&gt;kurtosis&lt;/sub&gt; (-0.418), LEXgrdBA1&lt;sub&gt;skewness&lt;/sub&gt; (0.417), LEXgrdFSOOUA1&lt;sub&gt;std&lt;/sub&gt; (0.457), LEXgrdFSOOUC1&lt;sub&gt;std&lt;/sub&gt; (0.402), LEXgrdFMLA1&lt;sub&gt;var&lt;/sub&gt; (0.463), LEXgrdFA1&lt;sub&gt;20P&lt;/sub&gt; (-0.56), LEXgrdFA2&lt;sub&gt;q1&lt;/sub&gt; (-0.531), LEXgrdFB2&lt;sub&gt;10P&lt;/sub&gt; (-0.441), LEXgrdFC1&lt;sub&gt;20P&lt;/sub&gt; (-0.403), LEXgrdFA2&lt;sub&gt;10P&lt;/sub&gt; (-0.497), LEXgrdFB1&lt;sub&gt;10P&lt;/sub&gt; (-0.45), LEXgrdFSOOUA1&lt;sub&gt;var&lt;/sub&gt; (0.457), LEXgrdFB2&lt;sub&gt;20P&lt;/sub&gt; (-0.483), LEXgrdFB1&lt;sub&gt;20P&lt;/sub&gt; (-0.497), LEXgrdFB2&lt;sub&gt;q1&lt;/sub&gt; (-0.485), LEXgrdFMLA1&lt;sub&gt;10P&lt;/sub&gt; (-0.411)</td>
</tr>
<tr>
<td><strong>Lexical diversity</strong></td>
<td>LEXdvrWLR&lt;sub&gt;avg&lt;/sub&gt; (0.545), LEXdvrFSS&lt;sub&gt;avg&lt;/sub&gt; (0.559), LEXdvrWSS&lt;sub&gt;avg&lt;/sub&gt; (0.587), LEXdvrFSR&lt;sub&gt;avg&lt;/sub&gt; (0.559), LEXdvrVLRW&lt;sub&gt;avg&lt;/sub&gt; (0.433), LEXdvrWLC&lt;sub&gt;avg&lt;/sub&gt; (0.545), LEXdvrNSR&lt;sub&gt;avg&lt;/sub&gt; (0.54), LEXdvrNSC&lt;sub&gt;avg&lt;/sub&gt; (0.478), LEXdvrVLRA&lt;sub&gt;avg&lt;/sub&gt; (0.478), LEXdvrVSR&lt;sub&gt;avg&lt;/sub&gt; (0.587), LEXdvrVLCA&lt;sub&gt;avg&lt;/sub&gt; (0.482), LEXdvrVSR&lt;sub&gt;avg&lt;/sub&gt; (0.558), LEXdvrVLR&lt;sub&gt;avg&lt;/sub&gt; (0.558), LEXdvrVSR&lt;sub&gt;avg&lt;/sub&gt; (0.449), LEXdvrVSR&lt;sub&gt;avg&lt;/sub&gt; (0.559), LEXdvrFSL&lt;sub&gt;avg&lt;/sub&gt; (0.558), LEXdvrVSL&lt;sub&gt;avg&lt;/sub&gt; (0.544), LEXdvrVSL&lt;sub&gt;avg&lt;/sub&gt; (0.544), LEXdvrVSSW&lt;sub&gt;avg&lt;/sub&gt; (0.449), LEXdvrNSS&lt;sub&gt;avg&lt;/sub&gt; (0.54), LEXdvrVSR&lt;sub&gt;avg&lt;/sub&gt; (0.482)</td>
</tr>
<tr>
<td><strong>Lexical errors</strong></td>
<td>ERRAllSUM&lt;sub&gt;avg&lt;/sub&gt; (-0.771)</td>
</tr>
<tr>
<td><strong>Lexical frequency</strong></td>
<td>LEXfrqNL&lt;sub&gt;20P&lt;/sub&gt; (0.446), LEXfrqLNS&lt;sub&gt;q1&lt;/sub&gt; (0.46), LEXfrqFCL&lt;sub&gt;20P&lt;/sub&gt; (0.448), LEXfrqFNL&lt;sub&gt;20P&lt;/sub&gt; (0.41), LEXfrqLC&lt;sub&gt;q1&lt;/sub&gt; (0.496), LEXfrqFCL&lt;sub&gt;q1&lt;/sub&gt; (0.464), LEXfrqLC&lt;sub&gt;20P&lt;/sub&gt; (0.488)</td>
</tr>
<tr>
<td><strong>Lexical sophistication</strong></td>
<td>LEXsopFK&lt;sub&gt;1&lt;sub&gt;var&lt;/sub&gt; (0.419), LEXsopLWK&lt;sub&gt;1&lt;sub&gt;avg&lt;/sub&gt; (-0.443), LEXsopLWK&lt;sub&gt;1&lt;sub&gt;skewness (0.417), LEXsopLWK&lt;sub&gt;1&lt;sub&gt;var&lt;/sub&gt; (0.427), LEXsopLWK&lt;sub&gt;1&lt;sub&gt;std (0.427), LEXsopFK&lt;sub&gt;1&lt;sub&gt;avg (-0.429), LEXsopFK&lt;sub&gt;1&lt;sub&gt;std (0.419)</td>
</tr>
<tr>
<td><strong>LexMorphology features</strong></td>
<td>LEXafxA&lt;sub&gt;avg&lt;/sub&gt; (0.442)</td>
</tr>
<tr>
<td><strong>Orthographic neighbors</strong></td>
<td>LEXnghNUM&lt;sub&gt;10P&lt;/sub&gt; (-0.448), LEXnghPHO&lt;sub&gt;qr&lt;/sub&gt; (0.426), LEXnghPHO&lt;sub&gt;0P&lt;/sub&gt; (0.417), LEXnghNUM&lt;sub&gt;20P&lt;/sub&gt; (-0.526), LEXnghNUMF&lt;sub&gt;30P&lt;/sub&gt; (-0.436), LEXnghFRQ&lt;sub&gt;q1&lt;/sub&gt; (-0.431), LEXnghPHO&lt;sub&gt;std&lt;/sub&gt; (0.432), LEXnghPHO&lt;sub&gt;q3&lt;/sub&gt; (0.425), LEXnghORT&lt;sub&gt;qr&lt;/sub&gt; (0.504), LEXnghAVG&lt;sub&gt;20P&lt;/sub&gt; (-0.42), LEXnghORT&lt;sub&gt;80P&lt;/sub&gt; (0.526), LEXnghORT&lt;sub&gt;0P&lt;/sub&gt; (0.492), LEXnghORT&lt;sub&gt;avg&lt;/sub&gt; (0.519), LEXnghNUMF&lt;sub&gt;10P&lt;/sub&gt; (-0.436), LEXnghPHO&lt;sub&gt;avg&lt;/sub&gt; (0.481), LEXnghNUM&lt;sub&gt;q1&lt;/sub&gt; (-0.515), LEXnghPHO&lt;sub&gt;max&lt;/sub&gt; (0.472)</td>
</tr>
<tr>
<td><strong>Word length</strong></td>
<td>LENwrdSYL&lt;sub&gt;80P&lt;/sub&gt; (0.416), LENwrdSYL&lt;sub&gt;q3&lt;/sub&gt; (0.449), LENwrdLETTERS&lt;sub&gt;var&lt;/sub&gt; (0.415), LENwrdSYL&lt;sub&gt;max&lt;/sub&gt; (0.468), LENwrdLETTERS&lt;sub&gt;sd&lt;/sub&gt; (0.406), LENwrdSYL&lt;sub&gt;dolch&lt;/sub&gt; (0.414), LENwrdSYL&lt;sub&gt;std&lt;/sub&gt; (0.517), LENwrdSYL&lt;sub&gt;avg&lt;/sub&gt; (0.558)</td>
</tr>
<tr>
<td><strong>Tense features</strong></td>
<td>SYNsFINDP&lt;sub&gt;avg&lt;/sub&gt; (-0.422)</td>
</tr>
<tr>
<td><strong>Text likelihood</strong></td>
<td>DISlkVSM&lt;sub&gt;kurtosis&lt;/sub&gt; (0.452), DISlkVSM&lt;sub&gt;avg&lt;/sub&gt; (-0.411), DISlkWLM&lt;sub&gt;q1&lt;/sub&gt; (-0.438), DISlkF&lt;sub&gt;20P&lt;/sub&gt; (-0.41), DISlkVLM&lt;sub&gt;min&lt;/sub&gt; (-0.424), DISlkWLM&lt;sub&gt;avg&lt;/sub&gt; (-0.469), DISlkF&lt;sub&gt;median&lt;/sub&gt; (0.401), DISlkNSM&lt;sub&gt;kurtosis&lt;/sub&gt; (0.401), DISlkWSM&lt;sub&gt;20P&lt;/sub&gt; (-0.48), DISlkVLM&lt;sub&gt;avg&lt;/sub&gt; (-0.4), DISlkF&lt;sub&gt;skewness&lt;/sub&gt; (0.409), DISlkVLM&lt;sub&gt;20P&lt;/sub&gt; (-0.451), DISlkVLM&lt;sub&gt;avg&lt;/sub&gt; (-0.509), DISlkWLM&lt;sub&gt;10P&lt;/sub&gt; (-0.416), DISlkF&lt;sub&gt;skewness&lt;/sub&gt; (0.455), DISlkVLM&lt;sub&gt;20P&lt;/sub&gt; (-0.453), DISlkALM&lt;sub&gt;10P&lt;/sub&gt; (-0.424), DISlkASM&lt;sub&gt;20P&lt;/sub&gt; (-0.422), DISlkF&lt;sub&gt;sd&lt;/sub&gt; (0.438), DISlkVSM&lt;sub&gt;median&lt;/sub&gt; (-0.417), DISlkASM&lt;sub&gt;min&lt;/sub&gt; (-0.406), DISlkWLM&lt;sub&gt;median&lt;/sub&gt; (-0.419), DISlkWLM&lt;sub&gt;sd&lt;/sub&gt; (0.4), DISlkVSM&lt;sub&gt;min&lt;/sub&gt; (-0.424), DISlkALM&lt;sub&gt;20P&lt;/sub&gt; (-0.402), DISlkWSM&lt;sub&gt;median&lt;/sub&gt; (-0.455)</td>
</tr>
</tbody>
</table>

Table 9: Correlation between features and CEFR level grouped by linguistic variable family. For the name of the features, see [https://cental.uclouvain.be/fabra/docs.html](https://cental.uclouvain.be/fabra/docs.html).