PADDLe: a Platform to Identify Complex Words for Learners of French as a Foreign Language (FFL)

Camille Pirali, Thomas François, Nuria Gala

Proceedings of the 2nd Workshop on Tools and Resources for REAding DIfficulties (READI 2022), pages 46–53
Marseille, 20-25 June 2022

PADDLe: a Platform to Identify Complex Words for Learners of French as a Foreign Language (FFL)

Camille Pirali[1], Thomas François[1,2], Núria Gala[3]
[1]Université catholique de Louvain, Belgium [2]CENTAL, IL&C
[3]Aix-Marseille Univ., CNRS, Laboratoire Parole et Langage (UMR 7309), France
camille.pirali@student.uclouvain.be, thomas.francois@uclouvain.be, nuria.gala@univ-amu.fr

Abstract

Annotations of word difficulty by readers provide invaluable insights into lexical complexity. Yet, there is currently a paucity of tools allowing researchers to gather such annotations in an adaptable and simple manner. This article presents PADDLe, an online platform aiming to fill that gap and designed to encourage best practices when collecting difficulty judgements. Studies crafted using the tool ask users to provide a selection of demographic information, then to annotate a certain number of texts and answer multiple-choice comprehension questions after each text. Researchers are encouraged to use a multi-level annotation scheme, to avoid the drawbacks of binary complexity annotations. Once a study is launched, its results are summarised in a visual representation accessible both to researchers and teachers, and can be downloaded in .csv format. Some findings of a pilot study designed with the tool are also provided in the article, to give an idea of the types of research questions it allows to answer.

Keywords: Text Simplification, Complex Word Identification, Lexical Difficulty, Lexical Complexity Prediction, Annotation Tool

1. Introduction

The importance of reading for language development, whether in an L1 or an L2, has been argued many times. However, in order for incidental learning of new vocabulary through reading to take place, it is necessary for the reader to already be familiar with the majority of the words they encounter (Huckin and Coady, 1999; Coady, 1996). Presenting readers with texts of an adequate difficulty level is thus essential to foster their reading skills and vocabulary development. This can be achieved either by comparing reading materials and choosing one of the desired level, or by simplifying elements of a text that are too complex. Both cases require to identify potential sources of difficulty for readers, notably on the lexical level.

Predicting how difficult a word will be for a reader requires large amounts of data, which should ideally be collected directly from the target population. Italian-speaking learners of French, for instance, are likely to struggle with different aspects of the language than Japanese speakers, who in turn will not have the same needs as French-speaking readers with dyslexia. Despite this fact, most of the literature devoted to predicting lexical complexity on the basis of difficulty annotations disregards demographic information and produces reader-independent measures of complexity. This one-size-fits-all approach is a first issue that we wish to address in this article.

A second issue is that there is a lack of tools and resources to collect such annotations of lexical difficulty. Indeed, researchers are typically faced with two options: they can either use crowdsourcing websites and create a batch of Human Intelligence Tasks (HITs) to be completed by workers, or create a custom-made platform from scratch. Both of these approaches present shortcomings that can be prohibitive: the first option tends to be expensive and may impose unwanted constraints on the format of the study, while the second requires web programming knowledge and can be very time-consuming. This is a shame, as it may render such studies inaccessible for some, despite them providing valuable insights for the scientific community.

The tool presented in this paper aims to make the process of collecting annotations simpler and accessible for other languages than English, as well as to encourage researchers to collect and account for demographic data. Designed primarily in order to analyse lexical difficulty for learners of French as a foreign language (FFL), it could easily be adapted to different target groups as well. Moreover, the tool strives to involve foreign language teachers in the data collection process, by allowing them to view their students' answers in real time and gain insights into the needs of their class.

The following section (2) will give an overview of previous methodologies employed when collecting similar data, in order to define key features that need to be taken into account. Section 3 will then describe the online platform PADDLe, highlighting the ways it responds to those observations and giving a few pointers on possible use cases. In Section 4, some results from a pilot study conducted through the platform will be presented. Finally, concluding remarks and some future areas of improvement will be proposed in Section 5.

2. Related Work

The task of identifying words which might pose a problem to readers has been referred to as Complex Word Identification (CWI) or Lexical Complexity Prediction

(LCP), depending on whether complexity is conceptualised on a binary or on a continuous scale.

2.1. Complexity Annotation Datasets

The success of a model attempting to predict lexical complexity is impacted by its architecture and the relevance of the selected features, but also by the quality of the data with which it is trained. This was made evident during the 2016 SemEval workshop (Paetzold and Specia, 2016), when the shortcomings of the dataset provided to participating teams for the CWI shared task were such that all teams performed rather poorly (Shardlow et al., 2021b).

The dataset collected for the shared task contained sentences extracted from corpora based on the standard and simplified versions of Wikipedia. Those sentences were annotated by non-native speakers of English, who were asked to assign a binary complexity label to each lexical word of a given sentence. A value of 1 indicated that the annotator could not understand the target word, regardless of whether they understood the meaning of the sentence as a whole. Sentences destined to make up the training set were annotated by 20 people each, while those forming the test set only received one annotation. Furthermore, the corpus was split in a rather unconventional way, with the test set being over forty times larger than the training set (Paetzold and Specia, 2016). This contributed to the complexity of the task according to Shardlow et al. (2021b), and is probably largely responsible for the poor results obtained by the submitted systems.

To build their predictive models, participating teams were presented with two versions of the training set. One version provided all individual annotations for each word, and was used by some teams to fine-tune their model. The other attributed a single tag to each word based on whether at least one reader had found it difficult (Paetzold and Specia, 2016). As a result, a word being marked as complex by only one of the annotators was considered just as complex as another to which all 20 annotators attributed a score of 1. Moreover, as pointed out by Shardlow et al. (2021b), binary complexity judgements rely on an arbitrary threshold decided upon by each annotator. A value of 1 in the training set might therefore have represented very different levels of complexity, which added to the overall difficulty of the challenge.

The subsequent edition of the task, organised in 2018, refined the collection and presentation of the data, which seems to have had a positive impact on the performance of submitted systems (Yimam et al., 2018). The second CWI shared task made use of a multilingual corpus, with languages being represented either in both the training and the test set (English, German, Spanish), or only in the test set (French). Annotations were collected through crowdsourcing, using Amazon Mechanical Turk (MTurk). This time, the data were split so that there would be a larger amount of training sentences than test sentences, and words in the test set received several annotations. Complexity was once again represented as a binary value, with a threshold of only one annotation required for a word to be considered complex. Interestingly, however, participants also received probabilistic values based on the proportion of annotators who did not understand a given word (Yimam et al., 2018). Unfortunately, such a probabilistic annotation system was not enough to make up for the shortcomings of binary annotations, as suggested by Shardlow et al. (2021b), who observed that a value of 0.5 only meant that a word was found complex by half of the annotators - and thus simple by the other half. As such, no direct conclusions could be drawn about its level of complexity.

The organisers of the shared task made several other methodological decisions that differed from the previous edition. While in 2016 the words to be annotated were predefined and presented in a sentence, this time annotators were given a paragraph of five to ten lines in which they were free to select up to ten complex items. This constraint might have had an impact on how complete the data were: indeed, it is possible that annotators sometimes had to make a choice when they had identified more than ten words they thought were complex. A second difference with the first edition of the task was that annotators were asked to identify complex multi-word expressions (MWEs) as well as complex words. This, combined with the fact that words were not preselected, might have impacted the data negatively as well. Indeed, Gooding and Kochmar (2018) reported that certain sequences of words were interpreted as single words by some of the annotators, and as MWEs by others. As for the annotators themselves, they were no longer non-native speakers selecting complex words based on their own understanding of them, but a mix of natives and non-natives who were asked to identify items that could be difficult for learners or people with a reading impairment (Yimam et al., 2018). This is an important distinction to make, as it is likely that some annotators selected words that they themselves understood, but assumed other people might not. The predictions obtained from those annotations might therefore not be equally reliable for all target profiles, and perhaps especially so for readers with a learning or reading disability.

Based on the limitations of those two shared tasks, Shardlow et al. (2021b) formulated a list of guidelines for future CWI datasets. These guidelines are:

1. The annotations should be continuous rather than binary;
2. The items to be annotated should be presented in context;
3. Multiple instances of a same item should be included in the dataset;
4. Each item should receive several annotations;

5. Annotators should represent a variety of profiles in terms of fluency and background;

6. Texts included in the corpus should represent different genres;

7. Both single words and multi-word expressions should be considered in the annotation process.

2.2. Recent Refinements

It is with these recommendations in mind that Shardlow et al. (2021a) compiled their own dataset for the 2021 shared task on Lexical Complexity Prediction. Similarly to Yimam et al. (2018), they collected annotations through crowdsourcing, using the Figure Eight (previously Crowdflower, now Appen) and Amazon Mechanical Turk platforms. They asked annotators to label the complexity of a word using a 5-point Likert scale, and took the mean of all annotations for an item as its gold-standard complexity. This system allowed them to obtain continuous values ranging from 0 to 1, thus moving from a binary classification task (complex or not complex) towards a prediction task estimating how complex a given word is. The items to annotate were preselected and presented in context, and each token occurred at least twice. Teams were therefore encouraged to take context into account when predicting complexity. Finally, the texts used to produce the dataset were taken from three diverse genres, and multi-word expressions were considered alongside single words in the annotation process.

Systems submitted for the task obtained encouraging results, with the highest-ranking teams being very close together in score. This implies that different approaches succeeded in modelling the data almost equally well, which can be interpreted as being a testament to the dataset's quality just as much as to the ingenuity of the participating teams. The recommendations laid out by Shardlow et al. (2021b) thus seem to be genuinely helpful when compiling a dataset for Lexical Complexity Prediction. It is why the tool presented in this paper was devised in a way that would allow researchers to adhere to each of the guidelines when gathering lexical difficulty annotation data.

A fundamental specificity of the three tasks presented above is that they aim to obtain a singular complexity value (whether binary or continuous), conceptualised as being intrinsic to the word itself and not dependent on the reader. What this means is that they assume a word to be somewhat universally complex or non-complex because of the characteristics it possesses, such as its frequency or its length. Inter-rater variability is expected, but seen almost as "noise" in the data resulting from subjective judgements rather than as a phenomenon of interest in itself. By contrast, when building a Text Simplification tool with a specific public in mind, the focus is likely to benefit from being shifted to lexical difficulty (*i.e.* how difficult someone perceives a word to be, based on their particular language knowledge) instead of complexity. Indeed, people are likely to have different simplification needs based on characteristics such as their proficiency level, reading disability or mother tongue. As a result, trying to predict their needs from those of a heterogeneous group might not always yield satisfactory results.

This idea is confirmed by the very low mean inter-rater agreement obtained in the 2016 edition of the CWI task (Krippendorff's α of 0.244) (Paetzold and Specia, 2016). Similarly, Yimam et al. (2017) reported lower agreement scores between non-native speakers than between native speakers of English, most likely due to the fact that the first group is more diverse in terms of language background and proficiency level. Agreement between the two groups was also low, which according to the authors indicates that their simplification needs might differ. As for the third edition of the task, (Shardlow et al., 2021a) didn't report any inter-rater agreement scores.

Since the three tasks did not aim to predict complexity for different groups of readers, no demographic data were provided to the participating teams, even though they had been collected for the first two editions. Nevertheless, Paetzold and Specia (2016) observed that the number of words deemed complex by an annotator was correlated with their age as well as level of proficiency in English. This goes to show that, as suggested by Gooding and Kochmar (2018), including demographic information at both the annotation and the prediction steps should increase the performance of models.

In a study whose methodology was inspired by the shared tasks, Tack (2021) addressed their shortcoming by taking individual differences into account. Via a custom-made online reading interface, L2 learners of French were presented with a set of texts (as opposed to sentences or paragraphs) based on their fluency level, so that all readers of the same level would annotate the same texts. They were asked to highlight any word that they personally found difficult, hence providing annotations of a binary nature. Two trials were organised to gather data: one with a smaller pool of participants (n = 9) with diverse L1 backgrounds, and one with a much bigger pool of participants (n = 47) all sharing the same mother tongue. An inter-rater analysis carried out for all annotators in the first trial yielded very similar results to those of Paetzold and Specia (2016), with a Krippendorff's α of 0.26. Grouping the annotators by proficiency level did not seem to have a clear impact on the metric: A2 readers got an α of 0.23, and B1 readers one of 0.30. Interestingly, the agreement rate between participants in the second trial, once grouped by proficiency level, was much higher: between 0.36 (B1 level) and 0.51 (B2 level). These results suggest that annotators with a similar profile (same mother tongue, proficiency level, education level and age, in this case) tend to agree more in their difficulty judgements than annotators with diverse profiles, which confirms the value of including demographic information in a CWI dataset.

It also follows that reading aids targeting a specific profile, such as dyslexic readers, adults with low literacy or L2 learners with a specific L1, would be likely to benefit from gathering data directly from that target population.

2.3. Summary and goals of our study

This brief overview of previous approaches to CWI/LCP dataset collection has shown the process to be a complex one, requiring many methodological decisions to be made. As the quality of the dataset appears to have a strong impact on the performance of models trained on it, making the right choices is of critical importance. This is why we created a tool that makes it possible for users to almost effortlessly design their own data collection process, and that encourages them to follow the recommendations formulated above. This tool will be further described in the next section.

3. Presentation of the Tool

PADDLe (Plateforme d'Annotation De la Difficulté LExicale) is an online platform[1] hosted by CENTAL, whose aim is to make CWI data collection easier. It currently only supports French, but should include a variety of other languages in the future. It allows researchers to create highly customisable web-based reading tasks and download the data in an easily parsable .csv format.

The tool sets out to make following the guidelines proposed by Shardlow et al. (2021b) easy. Researchers are encouraged to define a continuous annotation scale and to include multi-word expressions, as well as to gather demographic information from participants (which would make them aware of how diverse their annotator pool is). The platform plans for words to be presented in context, and for several participants to annotate the same texts. Finally, researchers are free to add as many texts as they want, and can thus easily include several instances of a word as well as texts of various genres in their corpus.

3.1. Interest

PADDLe was conceived to offer an alternative to other online survey builders. It is completely free of use, customisable, designed specifically for CWI data collection and does not require any web development knowledge. It also allows teachers who ask their students to participate in a reading task to view the results of their class afterwards, to thank them for their contribution.

3.2. Functionalities and Options

The design decisions made when developing PADDLe were based on the conclusions drawn from the literature presented in section 2. The reading tasks created through the platform have the following format:

1. **Demographic form**: Participants answer a series of questions selected by the researchers;
2. **Text annotation**: Participants annotate a text by clicking on words and MWEs they find difficult, according to an annotation scale. The scale, as well as the boundaries of clickable units, are defined by the researchers.
3. **Reading comprehension questions**: Participants' global comprehension of the text is tested using multiple choice questions. Once they submit their answers, participants are given feedback on whether they answered correctly.

Step 2 and 3 are repeated as many times as decided by the researcher before the study ends. All three steps of the task can be customised as follows:

1. **Demographic form**: Researchers can select any of the following: participants' identifier (if they don't want the data to be anonymous), age, country of origin, education level, target language proficiency level, other languages known and proficiency level in each of those languages, time spent learning the target language in a non-native context / in a native context, learning or readings disabilities, and "other" (which gives them the option to add an open question).
2. **Text annotation**: The task can include as many texts as necessary, and researchers can decide whether annotators will read all texts or only a subset of them. In the second case, participants can be presented with a) a set of texts chosen at random, b) all texts of a pre-defined and randomly selected group or c) one text drawn at random from each predetermined group. For each text, researchers provide a title, the id to use in the .csv files and the text itself, which must be formatted as one word or punctuation sign by line. This allows multi-word expressions (or other groups of words that are to be annotated as a unit) to be defined, by simply grouping them on the same line. Punctuation is not made clickable for annotators. Researchers are also asked to provide an annotation scheme to be used in the reading task: each annotation level is given a colour and a label. Currently, the interface only allows to have between 2 and 5 levels (in addition to the "no annotation" level). This is to encourage users to choose a non-binary scale of annotation.
3. **Reading comprehension questions**: For each text, users are asked to provide between 1 and 6 comprehension questions, each with 2 to 5 possible answers. They must also indicate which of the answers is correct.

Other parts of the study are customisable as well, such as the consent form to be read by participants before beginning the task or the text presenting the study on its home page.

[1] It is available at this address.

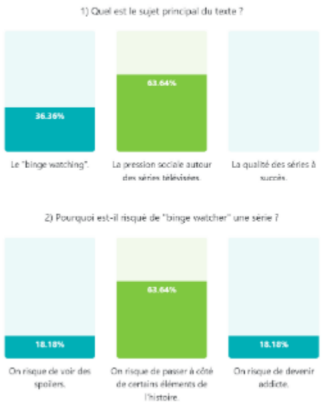

Figure 1: Visual representation of participants' answers to the comprehension questions.

Figure 2: Example of annotation.

3.3. Possible Uses

All the options presented above aim to make the generated studies as malleable as possible. Instead of whole texts, users of the platform could decide to only include paragraphs, or even sentences. Similarly, they could decide to ask for annotations at the phrase or at the sentence level, by grouping words in a way that suits their research purposes.

As a result, PADDLe could be used to answer a variety of research questions. One could for instance investigate the link between the proportion of words perceived as difficult and the quality of the general comprehension, based on the questions asked after each text. It could also serve to rate different simplifications of a text, in order to find the one readers understand best, or to compare a system's lexical complexity predictions with empirical difficulty judgements.

Outside of academic research, it could also prove an interesting tool for teachers who would like to pinpoint their students' difficulties. By asking all members of a class to annotate the same text and answer a few questions at the end, teachers could then refer to the visual representation of the results provided by the platform to immediately identify the words or aspects of the text found most difficult by the group. Figure 1 provides an example of said representation.

4. Pilot Study

To test the proper functioning and scientific interest of the interface before it could be used for larger research projects, a preliminary study was carried out in November 2021 with a small group of 16 L2 learners of French. It yielded some interesting first results, a selection of which will be presented in what follows.

The group was vastly homogeneous, as participants were all 18 year old students from Malaysia who shared the same mother tongue, all belonged to the B1 level in French and were currently following French classes at the Service Universitaire de Langues (SUL, Aix-Marseille university, France). The study used a total of three informative texts, and each participant was asked to annotate two of them - one seen by all participants (B1 level), and one randomly drawn from the remaining two texts (B2 level). As a result, the number of annotations is not balanced between the texts. Participants were asked to answer five comprehension questions after each text, in order to test their global understanding.

The annotation scale employed in the study, inspired by the Vocabulary Knowledge Scale (Wesche and Paribakht, 1996), a revised version of it (Sugiyama, 2017) and the scale used in the 2021 LCP task (Shardlow et al., 2021b), was the following:

0. **Easy** word, no annotation;
1. **Transparent** word (Unknown, but can guess the meaning in context);
2. **Vague** word (Unsure of the meaning);
3. **Opaque** word (Cannot understand the word at all).

Each difficulty level was represented by a colour and outline, indicated in a legend above the text. All words started on 0, and participants could click on a word to cyclically increase its difficulty level (once for "transparent", twice for "vague", three times for "opaque" and four to go back to "easy"). Figure 2 shows an example of what the annotation process looked like.

The scale aimed to capture increasing levels of difficulty, from familiar to entirely opaque. All words of the text could be annotated, regardless of their part of speech. As a result, it was expected that most words would receive a score of 0. By contrast with the scales mentioned above, ours only used one level for easy words. This was to avoid asking participants to annotate too many words, as including two different levels for familiar words would have required annotators to consider every single word in a text.

4.1. Overview of the Results

Initially, the number of annotators per text was as follows: 13 for text B1_A, 5 for B2_A and 10 for B2_B (a few participants only annotated one text instead of

two). However, we decided to discard any participation for which the total time spent on annotating a text and answering the questions was less than 60 seconds. As texts were between 432 and 564 words long and the average reading speed is about 250 words per minute, this seemed a more than reasonable threshold to enforce. Two annotations were thus discarded, from participants who spent 9.5 and 20.5 seconds on texts B1_A and B2_B respectively.

Table 1 provides some descriptive information about each text. On average, participants spent between 8 and 9 minutes on a task (annotation + questions), with a rather high level of variability between annotators. Every participant whose contribution was kept spent at least 2 minutes on a single task.

Texts	B1_A	B2_A	B2_B
Annotators	12	5	9
Number of words	432	564	448
% Easy	97.22	98.23	97.77
% Transparent	1.85	1.06	1.34
% Vague	0.93	0.35	0.22
% Opaque	0	0.35	0.67
Average task time (s)	521.9	557.4	508
Standard deviation (s)	180	156	234

Table 1: Descriptive statistics for annotated texts.

The percentages provided for each level of difficulty were calculated based on the mean score attributed to each word. As possible values ranged from 0 (no annotation) to 3 (opaque word), we chose the following thresholds for each level: 0-0.74 (easy), 0.75-1.49 (transparent), 1.5-2.24 (vague) and 2.25-3 (opaque). Those ranges were selected to separate the space into four equal parts, and were only used to provide an idea of the distribution of difficult words. There does not seem to be any noticeable difference between the three texts, although one could have expected the two B2 texts to have a lower proportion of easy words. However, the average difficulty value of words that received a label other than 0 from at least one participant is slightly higher in the more advanced texts: 0.59 (sd: 0.56) for B1_A, 0.76 (sd: 0.71) for B2_A, and 0.84 (sd: 0.76) for B2_B.

4.2. Inter-Rater Agreement

For each text, an inter-rater agreement analysis was carried out using Krippendorff's α for ordinal values (Krippendorff, 2011). The results are presented in table 2.

The values obtained in this study are significantly higher than the one reported for the 2016 edition of the CWI task (0.244, (Paetzold and Specia, 2016)), which could be due to the fact that the group of annotators was more homogeneous. However, major differences between the two experiments make comparison somewhat tricky. For one, the datasets used in both studies differ considerably in size: over 230,000 words were annotated by 400 participants in the 2016 task (Paetzold and Specia, 2016), for about 1,500 words and 16 annotators in the present pilot study. Although Krippendorff's α for ordinal data is less sensitive to the number of coders than other inter-rater agreement metrics (Antoine et al., 2014), such a difference in size cannot be overlooked. Moreover, the 2016 CWI dataset only included content words and was annotated in a binary manner, while this study made all words annotable and used a 4-point complexity scale.

Texts	B1_A	B2_A	B2_B
Raters	12	5	9
Krippendorff's α	0.45	0.50	0.57
Binary α	0.45	0.50	0.57

Table 2: Inter-rater agreement per text; comparatively high values show the benefits of taking demographic data into account.

By contrast, the study carried out by Tack (2021) is much more similar to this one and should therefore allow comparisons to be made: inter-rater agreement scores were computed for groups of 8 to 17 participants, and all words of the texts could be annotated. As mentioned in section 2, a similar score to the one reported by Paetzold and Specia (2016) was achieved by the group of 9 participants with diverse L1 backgrounds, while agreement rates ranging from 0.36 to 0.51 were obtained for the four groups made up of more homogeneous profiles. The agreement rates computed for the present pilot study confirm the finding that annotators with a similar profile produce congruent difficulty judgements. Converting the difficulty levels to binary labels (any value higher than 0 is set to 1) in our data to more closely match the settings of Tack's study had almost no impact on the agreement scores, as shown in table 2. This can be explained by the fact that having fewer possible labels makes agreement by chance between annotators more likely, and goes to show that, as suggested by (Antoine et al., 2014), the weighted nature of Krippendorff's α makes it less sensitive to the number of coding categories than other metrics.

4.3. Link Between Proportion of Difficult Words and Global Text Comprehension

The question of whether there was a correlation between the annotation provided by a participant and their performance when answering comprehension questions was explored using Spearman's rank correlation coefficient. This non-parametric measure was used as the data did not follow a normal distribution. The results are presented in table 3.

Two variables were tested for correlation with the num-

Texts	B1_A	B2_A	B2_B
Difficult x Correct	0.69	-0.11	0.13
Time x Correct	-0.08	0.72	-0.16

Table 3: Spearman's correlation tests (ρ).

ber of questions answered correctly by a participant: the proportion of words annotated as difficult (score of 1, 2 or 3) and the time spent on the task. The hypotheses were that 1) annotators who had found more words difficult would have a harder time answering the comprehension questions and 2) participants who spent more time on the annotation task would answer more questions correctly. In other words, we expected to find a negative correlation between percentage of annotated words and number of correct answers, and a positive correlation between time spent and number of correct answers.

Most of the results were inconclusive, and did not seem to support our hypotheses. The seemingly high positive correlation between the amount of time spent doing the task and the number of correct answers for text B2_A was not statistically significant (p = 0.086), probably due to the annotator pool being too small. Interestingly, a significant positive correlation was found between the proportion of difficult words and the number of correct answers for text B1_A ($p < 0.01$ with a one-tailed test going against our initial hypothesis). The same trend was found when aggregating all data, with a smaller but still significant positive correlation between the two variables (Spearman's ρ: 0.35, $p = 0.042$). This perhaps surprising result could be due to the fact that participants who completed the task more rigorously found a higher number of words to annotate. Indeed, a Spearman test between the time spent annotating a text and the proportion of words annotated as difficult found a small positive correlation between the two - however, it was not significant (Spearman's ρ: 0.23, $p = 0.134$).

It is worth noting that none of the annotators found more than 5% of the words of each text difficult (max: 4.63%). This implies that all participants were over the vocabulary coverage threshold of 95% that Laufer and Ravenhorst-Kalovski (2010) argue is required in order to understand a text properly. It is therefore likely that the positive correlation only holds true past a certain threshold, and that the trend would have been reversed had some participants found a higher proportion of the words of a given text difficult.

5. Conclusion and Future Work

The results presented in section 4 only scratched the surface as regards the types of exploratory analyses which could be undertaken using the tool. Data collected with it could also be fed to a predictive model, as was done during the shared tasks mentioned in section 2. Researchers could then make use of the demographic information that they are encouraged to gather in order to produce personalised predictions of difficulty.

The design of the tool aimed to address the shortcomings of previous approaches, namely the use of binary annotation data, the focus almost solely on English, and the low inter-rater agreement due to great heterogeneity in the pool of annotators. PADDLe is currently being used to gather data for a master's dissertation, which should further demonstrate the value of the interface.

Bibliographical References

Antoine, J.-Y., Villaneau, J., and Lefeuvre, A. (2014). Weighted Krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 550–559, Gothenburg, Sweden, April. Association for Computational Linguistics.

Coady, J., (1996). *L2 vocabulary acquisition through extensive reading*, page 225–237. Cambridge Applied Linguistics. Cambridge University Press.

Gooding, S. and Kochmar, E. (2018). CAMB at CWI shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194, New Orleans, Louisiana, June. Association for Computational Linguistics.

Huckin, T. and Coady, J. (1999). Incidental vocabulary acquisition in a second language: A review. *Studies in Second Language Acquisition*, 21:181 – 193.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability.

Laufer, B. and Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22:15–30.

Paetzold, G. and Specia, L. (2016). SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California, June. Association for Computational Linguistics.

Shardlow, M., Evans, R., Paetzold, G. H., and Zampieri, M. (2021a). SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online, August. Association for Computational Linguistics.

Shardlow, M., Evans, R., and Zampieri, M. (2021b). Predicting lexical complexity in english texts. *CoRR*.

Sugiyama, K. (2017). Analyse de la compétence lexicale dans la compréhension écrite des apprenants japonais en français. *Revue japonaise de didactique*

du français, numéro spécial : Actes du IVe Congrès régional de la Commission Asie-Pacifique, page 502 (12p.).

Tack, A. (2021). *Mark My Words! On the Automated Prediction of Lexical Difficulty for Foreign Language Readers*. Ph.D. thesis.

Wesche, M. B. and Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review-Revue Canadienne Des Langues Vivantes*, 53:13–40.

Yimam, S. M., Štajner, S., Riedl, M., and Biemann, C. (2017). CWIG3G2 - complex word identification task across three text genres and two user groups. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 401–407, Taipei, Taiwan, November. Asian Federation of Natural Language Processing.

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., and Zampieri, M. (2018). A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana, June. Association for Computational Linguistics.