Multilingual Resources for Lexical Complexity Prediction: A Review

Matthew Shardlow[1], Kai North[2], Marcos Zampieri[2]
[1]Manchester Metropolitan University, UK
[2]George Mason University, USA
m.shardlow@mmu.ac.uk

Abstract

Lexical complexity prediction is the NLP task aimed at using machine learning to predict the difficulty of a target word in context for a given user or user group to understand. Multiple datasets exist for lexical complexity prediction, many of which have been published recently in diverse languages. In this survey, we discuss nine recent datasets (2018-2024) all of which provide lexical complexity prediction annotations. Particularly, we identified eight languages (French, Spanish, Chinese, German, Russian, Japanese, Turkish and Portuguese) with at least one lexical complexity dataset. We do not consider the English datasets, which have already received significant treatment elsewhere in the literature. To survey these datasets, we use the recommendations of the Complex 2.0 Framework (Shardlow et al., 2022), identifying how the datasets differ along the following dimensions: annotation scale, context, multiple token instances, multiple token annotations, diverse annotators. We conclude with future research challenges arising from our survey of existing lexical complexity prediction datasets.

Keywords: Text Difficulty, Multilinguality, Lexical Complexity Prediction

1. Introduction

Estimating the complexity of words or multi-word expressions (MWE) to a reader is an important first step in automatic lexical simplification pipelines (North et al., 2023a). Lexical complexity is modeled using either Lexical Complexity Prediction (LCP) or Complex Word Identification (CWI). LCP is the task of assigning a value to a word which indicates how difficult that word will be for a reader (North et al., 2023b). This contrasts to CWI, which is the binary setting of identifying if a word requires simplification or not (Shardlow, 2013a; Paetzold and Specia, 2016; Zampieri et al., 2017). Two recent efforts to curate LCP resources were undertaken in recent shared tasks (Shardlow et al., 2021; Ermakova et al., 2022), resulting in the Complex 2.0 dataset (Shardlow et al., 2022) and the SimpleText 2022 Task 2 data. Whilst these resources focused solely on English, there have been significant efforts throughout the community to develop parallel resources in other languages (Pirali et al., 2022).

There is a great wealth of shared information between these resources and gathering them together into a single resource could benefit future multilingual complexity prediction applications. However, to unite these resources, we must understand the purposes of each resource and identify the parameters of their construction. The Complex 2.0 Framework provides seven recommendations for features of future LCP datasets. We have reproduced these with our interpretation below:

1. **Annotation Scale:** Whereas previous resources for identifying complex words had typically focussed on the binary case (complex or not), the Complex 2.0 Framework recommended the use of continuous annotations such as those resulting from aggregating over a Likert scale.
2. **Context:** The words to be assigned difficulty rankings were presented in context. Clearly context affects word sense, which affects difficulty, but also the surrounding context of a word may give some explanation or interpretation of that word.
3. **Multiple token instances:** The same word presented in different contexts gives rise to the opportunity to analyse the difficulty of a word across many occurrences.
4. **Multiple annotations per token:** Complexity is subjective and aggregating judgements from multiple diverse annotators will alleviate local subjective deviations.
5. **Diverse annotators:** Similarly, having a diverse group of annotators will help to give a representative sample of LCP annotations. The annotator pool may be targeted at a specific group (e.g., language learners, students, deaf people, et.) according to the intended application.
6. **Multiple genres:** Collecting texts from multiple genres allows for more diverse text types represented in the dataset.
7. **Multi-word expressions:** The inclusion of complexity predictions for MWEs as well as single word instances helps to give a more

Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024, pages 51–59

representative sample of the target language in the resulting LCP dataset.

In this paper, we survey the currently available LCP resources and analyse these through the lens of the recommendations given in the Complex 2.0 Framework. We note that (1) We identified nine suitable resources (listed in Table 1, CWI-18 appears three times, but is counted as a single resource) which we have focused this survey on. We only consider published available datasets in languages other than English, not work building on existing datasets. (2) the existing resources represent eight languages other than English. We do not include the English components of the three previous CWI/LCP shared task datasets (Paetzold and Specia, 2016; Shardlow et al., 2021; Yimam et al., 2018) or the English SimpleText 2022 Task 2 data (Ermakova et al., 2022) in our analysis. However for multilingual completeness, we do mention the French, Spanish and German components of the CWI-2018 Shared Task dataset.

The inclusion criteria for our resources were as follows:

- The resource was published since 2018.
- The resource provides complexity values of words at the level of single semantic units (i.e., not sentence or document level).
- The complexity values arise from annotation, as opposed to prediction or correlation to frequency.
- The language of the dataset was not English. We briefly discuss the existing English datasets, which have already been surveyed extensively below.

We provide an overview of the datasets we survey in Section 2, before progressing to a feature-based survey in line with the recommendations of the Complex 2.0 framework in Section 3.

2. Datasets Overview

2.1. CAS

Focused on technical terms in medical documents in French, Koptient and Grabar (2022) categorised terms from syntactic groups into 'understood', 'unsure' or 'not understood'. The authors gather lexical and syntactic features and train supervised learning algorithms to predict the reported difficulties of syntactic groups.

2.2. CWI18

Developed for a shared task at the BEA 2018 workshop, this dataset was developed in English, Spanish, French and German using Mechanical Turk to ask annotators to identify any words in a text that were complex. Each text was presented to multiple annotators, including native and non-native speakers, with complexity judgements applied to single words and spans. The final data was returned as both binary (did any annotator find the word complex) and continuous (how many annotators found the word complex).

2.3. VYTEDU-CW

A sample of Ecuadorian University students were asked to annotate texts from the VYTEDU corpus to indicate which words were difficult to understand. VYTEDU contains transcripts of educational videos in Spanish, which are suitable for university students. The authors provide some analysis of the complex words identified in the VYTEDU corpus, noting technical terms, sophisticated vocabulary, abbreviations, metaphor, unusual terms, verb-nominalisation and compound words as sources of complexity.

2.4. CLexIS^2

Students studying either Computer Systems or Software Engineering in Ecuador were asked to identify difficult words in transcripts of recorded lectures in Spanish from their courses using a custom annotation application. Complex words are later detected using an unsupervised and supervised approach.

2.5. LLCL

Lee and Yeung (2018) provide a study on the prediction of vocabulary knowledge for foreign language learners of Chinese. As a part of this study, they describe the annotation process of a dataset of Chinese words taken from the Lexical Lists for Chinese Learning in Hong Kong. Therein, they select 5 training sets and one test set which are labelled by language learners on a 5-point scale. These annotations focus on the word itself, without context presented.

2.6. RUBible

Texts from the Russian Synodal bible are annotated in a study closely replicating the work of (Shardlow et al., 2020). 931 words are presented across 3,364 contexts, which are then annotated on a 1–5 Likert Scale. The authors compare their results to the corresponding lexical complexity prediction data for English and also provide a linear regression demonstrating the ability to predict lexical complexity in Russian based on text features.

ID	Language	Reference
CAS	French	(Koptient and Grabar, 2022)
CWI18-FR	French	(Yimam et al., 2018)
CWI18-ES	Spanish	(Yimam et al., 2018)
VYTEDU-CW	Spanish	(Ortiz Zambrano et al., 2019)
CLexIS^2	Spanish	(Ortiz Zambrano and Montejo-Ráez, 2021)
LLCL	Chinese	(Lee and Yeung, 2018)
CWI18-DE	German	(Yimam et al., 2018)
RUBible	Russian	(Abramov and Ivanov, 2022)
JaLeCon	Japanese	(Ide et al., 2023)
CWITR	Turkish	(Ilgen and Biemann, 2023)
MultiLS-PT	Portuguese	(North et al., 2024)

Table 1: The datasets we consider for our survey. We have used the name given in the associated paper as the identifier, or the abbreviated name of the corpus that the LCP annotations are based on. In the case of the Russian dataset we have used the identifier RUBible as the texts are based on the Russian Synodal Bible.

2.7. JaLeCon

News and Government texts are provided to Native Japanese speakers as well as Chinese/Korean and other learners of Japanese for annotation on a 1-4 scale. Short word units and long word units are annotated with complexity values after word segmentation, which is necessary as Japanese does not support word boundaries. Baseline experiments show that a BERT-based system is effective for LCP in Japanese.

2.8. CWITR

Turkish language texts are annotated to identify complex words for readers using the binary setting. Annotations are collected for both complex words and phrases. Paragraph level texts are presented covering Wikipedia news, Wikipedia articles, news, novel summaries, and periodicals. All annotations were collected from native speakers of Turkish. In total 25 annotators provided complexity judgements over 13,837 instances.

2.9. MultiLS-PT

The MultiLS framework promotes a unified process for the tasks of lexical complexity prediction, substitution generation and binary comparative LCP. Brazilian Portuguese data has been collected for all tasks, but here we focus solely on the lexical complexity prediction data. This data is deliberately tied to the Complex 2.0 data, presenting 5,165 annotations across Bible, News and Biomedical texts.

2.10. English Datasets

Although not the main focus of this survey, there are English datasets available for complex word identification and lexical complexity prediction. The CW Corpus (Shardlow, 2013b) provided 731 instances of complex words mined from Simple Wikipedia edit histories. Later, related shared tasks Paetzold and Specia (2016) (Yimam et al., 2018) provided data for complex word identification. The Complex 2.0 (Shardlow et al., 2021) and SimpleText (Ermakova et al., 2022) corpora both provide English data for complexity prediction in Scientific texts (SimpleText) as well as religious and news (Complex2.0). Additionally, the work of Maddela and Xu (2018) provides word complexity data for 15000 words without contexts.

3. Literature Survey

3.1. Annotation Scale

The creators of LCP resources have used varied approaches to gather annotations. In all cases the resources that we have surveyed take the approach of identifying a target group and asking them a question about the difficulty of words in a text. The annotators are required to make a decision about the words, which may be a binary decision (is this word difficult or not difficult) (Ortiz Zambrano et al., 2019; Ortiz Zambrano and Montejo-Ráez, 2021; Ilgen and Biemann, 2023), or a graded decision on a Likert-scale (Koptient and Grabar, 2022; Lee and Yeung, 2018; Abramov and Ivanov, 2022; Ide et al., 2023; North et al., 2024). There is a subtle difference in the way that binary annotations or Likert-scale annotations are applied. In the binary setting, users are presented with an entire text and asked to mark any terms that they consider to be complex, with non-complex terms left unannotated. In the Likert-scale setting, annotators are presented with one or more tokens extracted from the text and asked to assign a rating based on a scale indicating difficulty. The annotator may choose to mark the word as an easy (low end of the scale) or difficult (high end of the scale) word. Binary annotations allow for a

much quicker annotation throughput as an annotator can return several annotations per sentence by simply highlighting all words they consider complex. Likert-scale annotations offer a more subtly graded degree of complexity. For instance, binary annotations ask 'Is the given word difficult to understand?', whereas Likert-scale annotations ask 'How difficult to understand is the given word?', returning an exact complexity value.

Binary annotations can be aggregated in two ways. Firstly, a researcher may choose to identify any word in a sentence as complex if at least one annotator considered it to be complex (Ortiz Zambrano et al., 2019; Ortiz Zambrano and Montejo-Ráez, 2021; Ilgen and Biemann, 2023). This returns a broad set of complex words without making a distinction between words that are considered complex by many or few annotators. To address this, probabilistic annotations (Yimam et al., 2018) aggregate the number of annotators that selected a word as complex in a binary setting. For example, in the CWI18 data 20 annotators identified complex words in each sentence. Each complex word has a probabilistic value derived as the number of annotators out of 20 that found the word to be complex.

Likert-scale data annotations are also collected from multiple annotators per instance and aggregated using 3 (Koptient and Grabar, 2022), 4 (Ide et al., 2023) or 5 (Lee and Yeung, 2018; Abramov and Ivanov, 2022; North et al., 2024) categories. Most examples of Likert-scale based datasets that we identified use simple mean averaging over the returned annotations to deliver a final complexity value following the Complex 2.0 framework (Abramov and Ivanov, 2022; Ide et al., 2023; North et al., 2024). A notable exception to this is CAS, which takes the most common annotation from their schema ('not understood','not sure' or unnanotated) as the overall label (Koptient and Grabar, 2022).

The LLCL dataset also reports a different construction technique which spans Likert-scale and binary protocols. In this dataset, the authors present a 5-point Likert-scale which is used for annotation by the target group (foreign language learners of Chinese). Annotators select a difficulty rating for each instance from 1 (Never seen the word before) to 5 (Absolutely know the word's meaning). The final dataset is then aggregated by considering any instances with an annotation of 5 as 'non-complex' and all others as 'complex' (Lee and Yeung, 2018).

3.2. Context

The mode of presentation of context at annotation time is an important decision to make in the construction of a LCP dataset. In the binary setting, the resources that we surveyed contain examples of tokens presented within a sentence (Ortiz Zambrano et al., 2019), paragraph (Ortiz Zambrano and Montejo-Ráez, 2021) and full document (Ilgen and Biemann, 2023). Allowing a reader to observe a full context allows them to explore the complexity of the word in context, taking into account both the specific word sense used as well as contextual factors such as clue words that may help to explain the difficult word. In the Likert-scale setting, we also observed examples of words presented within an entire document (Koptient and Grabar, 2022) as well as within a full sentence (Abramov and Ivanov, 2022; Ide et al., 2023; North et al., 2024).

The LLCL corpus (Lee and Yeung, 2018) only presents the word to annotators without context as the underlying corpus consists of a word list for foreign language learners of Chinese which are not presented within context. Datasets of words with lexical complexity annotations also exist for English (Maddela and Xu, 2018) and French CEFR levels (Pintard and François, 2020).

3.3. Multiple Instances of Each Token

This recommendation from the Complex 2.0 framework indicated that datasets for lexical complexity prediction should have several instances of the same token presented in-context. The perceived complexity of a word varies greatly depending on the presentation of the word in a sentence. Take, for example, the occurrence of the rare English word 'agog' in the following 3 examples from White (2017):

(1) They were **agog**.

(2) When the boy saw the sweets he was **agog** with anticipation.

(3) His talent [as a painter] is so enormous that you look at his surfaces with your mouth **agog** at the near-impossibility of it all.

In Example 1, it is very difficult to infer the meaning of the term. We can interpret that 'agog' is an emotion or sensation which can be held by a group of people but not much more. It is not clear from such a short context if this is negative or positive, abstract or concrete. Example 2 gives more context and a reader would correctly be able to interpret that 'agog' is related to the context term of anticipation and that it is the type of feeling a child may possess when seeing sweets. Even if the reader has never seen the term previously, they can infer the meaning from these contextual clues. Finally, in Example 3, a difficult word may appear within a context where the reader is led to incorrectly infer the meaning. In this case, a reader may be led to interpret 'agog' as a synonym of 'open', whereas in this case 'agog' is used to indicate eagerness or excitement.

In the datasets that we reviewed we found that all datasets which presented a context around the word also presented multiple instances of the same token. One particular variant to this approach is CAS, which uses syntactic groups to gather syntactically related terms for annotation (Koptient and Grabar, 2022).

3.4. Multiple Token Annotations

Lexical complexity is subjective (Shardlow, 2022). Two readers given the same text may identify different words as being complex. Moreover, two readers given the same word in the same context may assign a different complexity value on a Likert-scale. One factor that affects lexical complexity is L1 vs. L2 (Gooding et al., 2021; North and Zampieri, 2023), but this does not explain the full variation and more subtle factors such as education level, specialism and environmental factors are also likely to influence perceived complexity.

All the datasets we surveyed used multiple annotators to represent a variety of subjective opinions within the datasets. The degree of repeated annotations for the same instance varies widely across datasets with the CWI18 datasets reporting as few as 2 annotations per instance (Yimam et al., 2018) ranging up to 5-7 (Koptient and Grabar, 2022; Ortiz Zambrano and Montejo-Ráez, 2021; Ilgen and Biemann, 2023; Lee and Yeung, 2018) or even more than 10 (Abramov and Ivanov, 2022; Ide et al., 2023). More annotations per instance allows for a diverse range of subjective opinions to be represented and for the aggregation of these opinions to represent some normative value that can be useful for all annotators.

One strategy for collecting multiple annotations is to use crowdsourcing (Yimam et al., 2018; Ilgen and Biemann, 2023; North et al., 2024). Many resources that we surveyed do not report whether the annotators were paid or unpaid (Koptient and Grabar, 2022; Ortiz Zambrano et al., 2019; Ortiz Zambrano and Montejo-Ráez, 2021; Lee and Yeung, 2018; Ide et al., 2023). In these cases we assume that annotators were selected from populations that did not require remuneration (such as colleagues or students). Several authors report using Mechanical Turk, but do not report the amount paid per instance (Yimam et al., 2018; Abramov and Ivanov, 2022). 2 of the resources that we surveyed do report the degree of pay for the annotators, with RUBible paying 10 cents for a batch of 10 instances and MultiLS-PT reporting payment of 2 cents per instance.

3.5. Diverse Annotators

Annotators vary between multilingual datasets. Annotators have been either hand-selected or crowd-sourced and are representative of differing target demographics. Several datasets were developed to create LCP systems for second-language (L2) learners (Lee and Yeung, 2018) and have subsequently been annotated by individuals not native to the dataset's target language. Other datasets are developed solely for identifying complex words for first-language (L1) speakers (Ortiz Zambrano et al., 2019; Abramov and Ivanov, 2022). These datasets are annotated by individuals native to the predominant language of the dataset. However, other annotator variables are often controlled, including age, level of education, or reading disability. The following paragraphs discuss the merits and flaws of datasets that have (a) employed hand-selected versus crowd-sourced annotators, alongside (b) controlled influential annotator variables.

Several multilingual datasets hand-selected their annotators making them ideal for the creation of personalised LCP systems. CAS (Koptient and Grabar, 2022) hand-selected 9 French speaking annotators to rate the complexity of medical jargon for non-expert patients. By hand-selecting their annotators, (Koptient and Grabar, 2022) were able to control the level of prior familiarity annotators had with medical terminology improving the validity of their gold complexity labels. They only selected annotators with no self-reported medical knowledge, and asked annotators to not refer to online material, including dictionaries, for assessing word difficulty. VYTEDU-CW (Ortiz Zambrano et al., 2019) and CLexIS (Ortiz Zambrano and Montejo-Ráez, 2021) hand-selected university students in Ecuador to identify complex words spoken in Spanish. They likewise controlled annotator familiarity by presenting annotators with transcripts of recorded lectures that were on a subject-matter known but not overly familiar to the annotators. LLCL (Lee and Yeung, 2018) and JaLeCon (Ide et al., 2023) hand-selected 7 and 15 L2 learners of Chinese and Japanese respectively. Both datasets make reference to L2 proficiency frameworks, with Ide et al. (2023) having only recruited annotators with at least an intermediate level of L2 proficiency.

CWI18 (Yimam et al., 2018), CWITR (Ilgen and Biemann, 2023), and MultiLS-PT (North et al., 2024) crowd-sourced annotators using Amazon Mechanical Turk (MTurk), whereas RUBible (Abramov and Ivanov, 2022) crowd-sourced their annotators from Toloka. As such, each dataset was able to obtain a substantially greater number of annotators compared to those datasets that adopted hand-selection. The CWI18-FR and CWI18-ES datasets (Yimam et al., 2018) were annotated by 22 and 54 respectively, and were recruited from a variety of countries. CWITR (Ilgen and Biemann, 2023) hired 25 annotators located in Turkey, MultiLS-PT (North et al., 2024) selected 25 an-

notators from Brazil, and RUBible (Abramov and Ivanov, 2022) gathered 10 separate annotators from Russia, Ukraine, Belarus and Kazakhstan. However, only several of these datasets attempted to control language proficiency. The CWI datasets make a distinction between native and non-native speakers yet do not explain how this distinction has been made. CWITR (Ilgen and Biemann, 2023) enforced a language proficiency exam to record Turkish language proficiency. The remaining datasets were unable to collect information regarding mother tongue, number of languages known, or L2 proficiency. Past studies have shown that discrepancies in these variables between annotators results in differing perceptions of word difficulty (Maddela and Xu, 2018; North and Zampieri, 2023). Failure to control these variables is an obvious drawback which reduces the validity of crowd-sourced datasets. This is only compensated by their larger pool of annotators and overall generalisability.

3.6. Multiple Genres

Multilingual datasets differ in genre. Several datasets contain texts pertaining to a single genre (Koptient and Grabar, 2022; Yimam et al., 2018; Ortiz Zambrano et al., 2019; Abramov and Ivanov, 2022). Other datasets consist of multiple genres (Ide et al., 2023; Ilgen and Biemann, 2023; North et al., 2024). These genres include medical-related articles, educational materials, the Bible to news and Wikipedia extracts. These genres are typically believed to be of great importance. They relate to such topics as health literacy, education, or political awareness motivating their simplification for improved accessibility (North et al., 2023b). The following paragraphs detail the types of texts provided by the single and multi-genre datasets shown within Table 1 and summarise their uses.

Single genre datasets include CAS (Koptient and Grabar, 2022), the CWI18 datasets (Yimam et al., 2018), VYTEDU-CW (Ortiz Zambrano et al., 2019), CLexIS (Ortiz Zambrano and Montejo-Ráez, 2021), and RUBible (Abramov and Ivanov, 2022). CAS provides a corpus of 100 clinical reports annotated with complex words. These clinical reports summarise a patient's medical history, diagnosis and outcome. CWI18-FR, CWI18-ES, and CWI18-DE provide 2,251, 14,280, and 7,403 complex words in context taken from Wikipedia articles (Yimam et al., 2018). Wikipedia articles are a common source of texts for LCP researchers. Public edits to pre-existing articles were previously used to gather gold complex and simplified labels (Shardlow, 2013b). Later datasets, such as the CWI18 datasets, improved their validity by incorporating human annotation. VYTEDU-CW (Ortiz Zambrano et al., 2019) and CLexIS (Ortiz Zambrano and Montejo-Ráez, 2021) gathered educational material in the form of transcripts from university lectures. These datasets are unique as they provide instances that contain elements of spoken language. The Bible is another popular text for LCP researchers. RUBible contains 3,364 extracts parallel to those found within an English sister dataset, CompLex 2.0 (Shardlow et al., 2020). RUBible is therefore a perfect dataset for the investigation of cross-lingual transfer learning in regards to LCP.

Single genre datasets allow for model specialisation, whereas multi-genre datasets are used to report model performances across multiple domains. Models trained on several single genre datasets or one multi-genre dataset can be used to investigate the performances of unique training strategies, such as transfer learning between genres and in some instances, cross-lingual transfer learning.

3.7. Multi-word Expressions

Lexical complexity prediction can be applied both to single words and to multi-word expressions (defined as a contiguous set of tokens separated by white space, with a single well-known meaning). English datasets for complex word identification and lexical complexity have taken multi-word expressions into account (Yimam et al., 2018), (Shardlow et al., 2022). In this context, we treat multi-word expressions as single lexical units, which behave as words. We assume that a complexity judgement can be made regarding a multi-word expression in the same way that it can be made for a single word. Non-compositional multi-word expressions hold some semantic value that cannot be derived from the meaning of constituent words. E.g., a hot dog is not a type of dog and may not even be hot. Similarly, the complexity of a non-compositional multi-word expression may not be easily derivable from the complexities of its constituent words.

In our multilingual resources, we observed 3 instances of datasets which report solely on single-word lexical complexity (Ortiz Zambrano and Montejo-Ráez, 2021; Lee and Yeung, 2018; North et al., 2024). All other resources took MWEs into account. The idea of MWEs comes from the English language and the idea of single- vs multi-word units may not transfer easily to other languages. For example in a language such as German, there is a heavy degree of noun compounding, where spaces between words are omitted. These behave as multi-word expressions, but appear as single words. This is particularly apparent for Japanese, which mixes syllabic and logographic characters without word boundaries. The JaLeCon dataset provides annotations over Short Unit Words (SUWs) which correspond to one or two small lexical units. Multi-word expressions are identified as Long Unit Words (LUWs), which are also annotated for complexity. (Koptient and Grabar, 2022) use syntactic groups

to form token sequences that are then annotated for complexity. These may be single words, but are often several contiguous words under a single syntactic head.

4. Discussion

The most stark difference in the resources that we have surveyed is the question that is presented to the judges of lexical complexity. In the binary setting annotators are asked to identify any complex words (and often also phrases) in a text, whereas in the Likert-scale setting annotators are asked to return a judgement on a multi-point scale for a given word (usually) in a context. This gives rise to two very different forms of lexical complexity datasets. The former refers to words or phrases which have been identified as problematic by some user. the latter refers to words or phrases which have been assigned some value judgement according to their complexity. Researchers working with both types of data should bear in mind the difference between these protocols. A 0 (non-complex) label in the binary setting implies no user found this word to be complex, whereas a 0 label in the Likert-scale setting implies that every user indicated this word to be the least complex. Similarly a 1 (complex) label in the binary setting implies that at least 1 user (depending on the aggregation protocol used) found this instance to be complex, whereas a label of 1 in the Likert-scale setting implies that every user rated the word as the most difficult complexity level.

Additionally, it is worth considering that in the binary setting a user may be asked to identify any complex words or phrases in a text, whereas in the Likert-scale setting pre-identified words are presented. Both these processes may lead to biases in datasets (reflecting tokens selected by the annotators, or tokens selected by the researchers), which should be considered when making decisions about what is desired from the resulting dataset. For example, a researcher may want only examples of complex language in an LCP dataset, in which case they may select specific tokens according to some pre-identification protocol. Alternatively, a researcher may wish to have both low-complexity and high-complexity elements in a dataset in which case they may select tokens at random.

The datasets that we have identified cover 8 languages. Including 5 Indo-European languages (French, Spanish, German, Russian and Portuguese), 6 alphabetic languages (French, Spanish, German, Russian, Turkish and Portuguese), 2 Logographic languages (Chinese and Japanese), with Japanese also exhibiting Syllabary elements (Kana). Notable exceptions include south asian langauges (e.g., Hindi, Urdu, Sinhala, Bengali) and African languages as well as other low-resource languages.

The resources that we have surveyed present a variety of languages, but also text genres incorporating encyclopaedia text, medical texts, educational texts and religious texts. Systems trained for one language or genre may be more easily adaptable to future related languages and genres. This allows for the creation of generalisable models that are able to perform well on varying types of texts.

There is some variability in the protocols used for annotation. For example, the number of annotators per instance varies from 2 to 25. It is important for dataset providers to report on these statistics and to release appropriate metadata alongside the annotations to allow future users of lexical complexity prediction datasets to fully understand the meaning of the annotations. One particular source of variability is the use of native speakers, non-native speakers or language learners as annotators. It is likely that each group will have different complexity needs and will return different subjective lexical complexity judgements. Ongoing work on personalised lexical complexity (Gooding and Tragut, 2022) could benefit from varied datasets, if the appropriate metadata for target groups is maintained.

5. Future Research Challenges

The Complex 2.0 framework and MultiLS framework describe a pattern for future dataset creation for lexical complexity prediction resources and beyond. Future resources can follow the recommendations found in these works to deliver future datasets in diverse languages conforming to robust protocols followed by previous datasets. The MLSP shared task[1] is currently seeking to create a new dataset following the MultiLS framework for both lexical complexity prediction and lexical simplification. Future work to extend these datasets with additional languages, additional annotations in existing languages and additional text types will be beneficial to the community in generating new and interesting types of data for lexical complexity prediction in diverse lingual settings. We would particularly like the community to prioritise: (a) the development of LCP resources for widely spoken languages such as Mandarin Chinese, Hindi, Arabic, Bengali and beyond. (b) the inclusion of diverse language families beyond the heavy tendency to develop resources for Indo-European languages. (c) LCP resources for low-resourced languages.

[1] https://sites.google.com/view/mlsp-sharedtask-2024/home

References

Aleksei V. Abramov and Vladimir V. Ivanov. 2022. Collection and evaluation of lexical complexity data for russian language using crowdsourcing. *Russian Journal of Linguistics*, 26(2):409–425.

Liana Ermakova, Eric Sanjuan, Jaap Kamps, Stéphane Huet, Irina Ovchinnikova, Diana Nurbakova, Sílvia Araújo, Radia Hannachi, Elise Mathurin, and Patrice Bellot. 2022. Overview of the clef 2022 simpletext lab: Automatic simplification of scientific texts. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 470–494. Springer.

Sian Gooding, Ekaterina Kochmar, Seid Muhie Yimam, and Chris Biemann. 2021. Word complexity is in the eye of the beholder. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4439–4449, Online. Association for Computational Linguistics.

Sian Gooding and Manuel Tragut. 2022. One size does not fit all: The case for personalised word complexity models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 353–365, Seattle, United States. Association for Computational Linguistics.

Yusuke Ide, Masato Mita, Adam Nohejl, Hiroki Ouchi, and Taro Watanabe. 2023. Japanese lexical complexity for non-native readers: A new dataset. In *Proceedings of the Eighteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

Bahar Ilgen and Chris Biemann. 2023. Cwitr: A corpus for automatic complex word identification in turkish texts. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '22, page 157–163, New York, NY, USA. Association for Computing Machinery.

Anaïs Koptient and Natalia Grabar. 2022. Automatic detection of difficulty of French medical sequences in context. In *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*, pages 55–66, Marseille, France. European Language Resources Association.

John Lee and Chak Yan Yeung. 2018. Automatic prediction of vocabulary knowledge for learners of chinese as a foreign language. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–4. IEEE.

Mounica Maddela and Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760, Brussels, Belgium. Association for Computational Linguistics.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2023a. Deep learning approaches to lexical simplification: A survey. *arXiv preprint arXiv:2305.12000*.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2024. MultiLS: A Multi-task Lexical Simplification Framework. *arXiv preprint arXiv:2402.14972*.

Kai North and Marcos Zampieri. 2023. Features of Lexical Complexity: Insights from L1 and L2 Speakers. *Frontiers in Artificial Intelligence*, 6(1).

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023b. Lexical complexity prediction: An overview. *ACM Comput. Surv.*, 55(9).

Jenny Ortiz Zambrano, Arturo MontejoRáez, Katty Nancy Lino Castillo, Otto Rodrigo Gonzalez Mendoza, and Belkis Chiquinquirá Cañizales Perdomo. 2019. Vytedu-cw: Difficult words as a barrier in the reading comprehension of university students. In *The International Conference on Advances in Emerging Trends and Technologies*, pages 167–176. Springer.

Jenny A. Ortiz Zambrano and Arturo Montejo-Ráez. 2021. CLexIS2: A new corpus for complex word identification research in computing studies. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1075–1083, Held Online. INCOMA Ltd.

Gustavo Paetzold and Lucia Specia. 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. Association for Computational Linguistics.

Alice Pintard and Thomas François. 2020. Combining expert knowledge with frequency information to infer CEFR levels for words. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 85–92, Marseille, France. European Language Resources Association.

Camille Pirali, Thomas François, and Núria Gala. 2022. PADDLe: a platform to identify complex

words for learners of French as a foreign language (FFL). In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 46–53, Marseille, France. European Language Resources Association.

Matthew Shardlow. 2013a. A comparison of techniques to automatically identify complex words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109, Sofia, Bulgaria. Association for Computational Linguistics.

Matthew Shardlow. 2013b. The CW corpus: A new resource for evaluating the identification of complex words. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 69–77, Sofia, Bulgaria. Association for Computational Linguistics.

Matthew Shardlow. 2022. Agree to disagree: Exploring subjectivity in lexical complexity. In *Proceedings of the 2nd Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI) within the 13th Language Resources and Evaluation Conference*, pages 9–16, Marseille, France. European Language Resources Association.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — a new corpus for lexical complexity prediction from Likert Scale data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 57–62, Marseille, France. European Language Resources Association.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Matthew Shardlow, Richard Evans, and Marcos Zampieri. 2022. Predicting lexical complexity in english texts: the complex 2.0 dataset. *Language Resources and Evaluation*, 56(4):1153–1194.

Murray White. 2017. Alex Janvier and the fine art of defiance. Toronto Star.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 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. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Gustavo Paetzold, and Lucia Specia. 2017. Complex word identification: Challenges in data annotation and system performance. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, pages 59–63, Taipei, Taiwan. Asian Federation of Natural Language Processing.