MLSP Shared Task @ BEA 2024

Annotation Guidelines - V1.0

These guidelines are intended for data providers for the proposed Shared Task at BEA 2024. Our intention is to collect a multilingual test set of lexical simplification pipeline operations covering lexical complexity prediction and substitution generation. We will direct participants to existing datasets for these tasks as training data, namely the CompLex data for Lexical Complexity prediction and the TSAR-2022 shared task data for the prediction of lexical simplifications in English. Datasets for these two tasks also exist for other languages, and we expect one of the main challenges in participation to be the adaptation of existing resources for the languages that we will provide.

To provide data for this task, you will need to follow these steps, which are detailed through the rest of this document:

- **Data Preparation**
- **Annotator selection**
- **Data Annotation for Lexical Complexity**
- **Data Annotation for Lexical Simplification**
- **Data Aggregation**
- **Post-hoc human evaluation**

We have also included information on **Ethics** and **Publication**, which can be found at the end of this document.

Data Preparation

You should select 200 words in your target language, for a relevant definition of word in your language. Words should typically be single lexical units, not covering multiple semantic boundaries. For example, in English we will not select multi-word expressions, but focus on single words. The words should be selected using appropriate heuristic criteria for your given language to ensure that the selected words are sufficiently difficult to warrant lexical complexity annotation, and particularly to ensure that annotators will be able to find some simpler substitutions for the word in context. To help guide selection, we will provide a sample list of 200 words in English. You may choose to translate part of this list, or simply use it as a guide to understand the type and distribution of words that you should select.

Once the words have been selected, you should identify 200 contexts in your target language, each context should contain one of the target words.

The contexts should be selected from a readily available source in your language that is related to educational settings. The source must be released under a license that allows further redistribution of the text to ensure that we are able to release the annotations with the original texts. This may be news articles, story texts, or encyclopedic texts, but you may choose other genres if they are the most suitable for your language. The source texts should be written by native speakers, and should not be the result of automated or manual translation. For each context, you should also select an additional 2 words in that context for annotation. You may choose any words that are representative of the informational content, according to heuristics that will allow you to identify interesting words for annotation in your examples.

The requirement to select 200 contexts, with 3 words per context gives rise to 600 instances in total. We are asking for a smaller set of common contexts to reduce the effect of context complexity. I.e., other complex words in a context might affect the perceived complexity of the target words. By providing common annotations on a smaller set of contexts we will have more commonality and less effects from context variation. This is a soft requirement on the data and you may choose to interpret this as you wish. For example, there may be some languages where the contributor already has 600 unique targets selected across 600 contexts. In these cases you may wish to disregard the common contexts requirement and provide 600 unique contexts.

An example is given below, with the selected target words highlighted in bold text:

Folly is set in **great dignity.**

Note that the highlighted words: ‘Folly’, ‘great’ and ‘dignity’ all bear semantic content. The remaining words (the copula ‘is’ and the preposition ‘in’) are short words that do not have much influence on the overall meaning of the text. Particularly, it would be hard to find substitutions for these words. A good rule of thumb (for languages similar to English) could be to select longer words that are nouns, verbs or adjectives, but you should consider what selection criteria will be appropriate for the language for which you are providing data. We expect most words that are selected to be sufficiently difficult that annotators will be able to find simpler substitutions for them.

Annotator Selection

We request that you provide at least 10 annotations per instance for this task. You may choose to identify known annotators, or recruit the annotators through an online platform such as mechanical turk or Prolific. Native speakers are preferred for annotation, but including some non-native speakers may lead to interesting opportunities for analysis at a later point. If there is any other information from the annotators that you consider relevant to keep, please include it in the metadata.You should record the following elements of metadata for each annotator:

- the number of years the annotator has spent in education
- whether or not they are a native speaker of the language that is being annotated.
- Age

- Typical number of hours they spend reading per week
- First Language
- Number of languages they speak

When selecting your annotators, you should consider and record the ‘target group’ of your annotations. This will depend on the annotator pool that you have access to. For example, if your annotators are selected from students at your university then the target group could be considered to be students in a higher educational setting. If your annotators are drawn from a language learning school, then you should record this as the annotations will be reflective of the needs of language learners.

You may choose as many annotators as you require. Please record demographic information for each annotator as above. Depending on the target group, you may choose the same annotators to perform both LS and LCP, or you may choose different groups for each task. For example, if you are targeting language learners, you may wish to ask them to provide LCP annotations and ask their teachers to provide LS annotations. Please record demographic information for both groups. Whilst each instance should have at least 10 annotations, it is not necessary to show every annotator every instance. For example, you may have 20 annotators and show 50% of the instances to each annotator. Each instance would still receive 10 annotations in total. Please keep a record of the metadata for each individual annotator. You should also record the annotations provided by each individual annotator in an unaggregated format to allow the calculation of inter-annotator agreement.

Data Annotation for Lexical Complexity

Annotations for lexical complexity should be done using a 5-point Likert scale, with the following points translated appropriately:

1. Very Easy - Words which are very familiar to you
2. Easy - Words which are mostly familiar to you
3. Neutral - When the word is neither difficult or easy
4. Difficult - Words which you are unclear of the meaning, but may be able to infer from the context
5. Very Difficult - Words that you have never seen before, or are very unclear

Each instance should be presented to the annotator with the full context. The annotators should be asked to provide an independent judgment for each of the three highlighted words per context. You should record the judgment (1,2,3,4 or 5) given to each word by each annotator. These will be processed to provide a single dataset.

Note that in the Likert scale descriptors we have suggested a subjective wording of the task. I.e., how complex is this word for *you*. You may wish to rephrase this when translating the

descriptors in accordance with your intended target group. For example, if your annotations are intended to be suitable for children, you might ask ‘how complex is this word *for a typical child*?’

You should also consider quality control of the annotations through manual review. Particularly if you are using crowd-sourcing such as through Amazon Mechanical Turk. In the ideal case, you should review a sample of the work of every annotator to ensure that they have understood the task and that the responses that are given are appropriate (i.e., not just selecting very-easy or very-hard for every instance, or random-clicking).

Data Annotation for Lexical Simplification

For each highlighted word, annotators should provide a minimum of 1 and a maximum of 3 words that could be used to simplify it in the given context. The words should be selected to ensure (a) that the meaning of the original word and the overall context is preserved and (b) that the word that is returned is a genuinely easier to understand alternative. The target word should be presented in its original context and the annotators should be instructed to provide replacement words that fit correctly within the given context. If more than one replacement is given, there is no need for the annotators to provide an ordering of the best-worst fit. All replacements that are given should satisfy the criteria of meaning-preservation and simplicity.

Some of the selected words may not be reasonably simplifiable in the contexts that they are presented in. I.e., a word may already be sufficiently simple, or despite being complex there may be no simpler alternative to a word. In these cases the annotators should write the original word or leave the field blank (indicating that the original word is the simplest word that could fit in this context. In the vast majority of cases, annotators should be able to find a simpler alternative to the given target words.

Again, you should provide some quality control through manual checking of the results of each annotator. Some annotators may find this task more difficult than others and there is likely to be some subjective variation in the number of replacements and the simplicity level of suggested replacements. A quality control element to identify is the frequency with which annotators have been unable to find a simplification. If this is unusually high compared to other annotators for the language then this may be cause for further review of the annotator’s results.

For some languages, a substitution may cause issues regarding the agreement with surrounding words (e.g. a masculine noun replaced by a feminine substitution will require to revise the gender of its related adjectives or determiners). We decided to treat morphological adaptation as a separate task that is left aside, so annotators should be informed that they may propose substitutions that do not strictly fit in the grammatical context regarding the agreement.

Data Aggregation + Return

Once all 600 instances (200 contexts, with 3 targets each) have been annotated, the data should be collected and returned in a tab-separated format with the headers below. For the LCP data, the mean average of the Likert scale points should be returned. For the LS data, the suggestions should be ordered by their frequency of suggestion following the same format as the TSAR-2022 dataset.

<Token> <Context> <Avg LCP>
<Most_frequent_suggestion>...<Most_frequent_suggestion><Second_most_frequent_suggestion>...<Second_most_frequent_suggestion><Third_most_frequent_suggestion> ….

You should also record and return the following metadata instances:

- Number of annotators per instance (average if not the same across all instances)
- Annotator demographics (averaged over all annotators)
- Intended target audience
- Text genre(s) represented in contexts

Additionally to the aggregated data, there will be interesting unaggregated data as a result of the annotation. We do not plan to release annotator-level unaggregated data to our task participants, so no specification is given in this document for that data. There may be opportunities to perform inter- and intra-lingual analyses of the unaggregated data outside the context of the shared task. We leave this to data providers to consider.

Post-hoc human evaluation

For each language we will select a sample of the outputs for each of the top 5 performing systems. We will select a common sample of the annotations per system. These will be judged for **simplicity**, **fluency** and **meaning preservation** on a scale of 1-5. We will ask task participants to help with the manual judgments for their target language(s). We expect to have a minimum of 2 judges per language, but we will try to recruit more than this. Data providers may also serve as judges for their provided languages.

Ethical Considerations

When conducting your annotation studies, please seek any necessary ethical approval from your institution. You should record any payment made to annotators, and the payments should constitute a wage that is equivalent to, or above, minimum wage in the annotators' country of residence. Please inform annotators that all information data that is gathered, including demographic data, will be stored in an anonymous format and distributed for research purposes. Annotators should have a chance to view the specifications of the task, including how data will be shared prior to deciding whether to complete the task or not.