Introducing MultiLS-IT: A Dataset for Lexical Simplification in Italian

Laura Occhipinti[1]

[1]University of Bologna, Italy

Abstract

Lexical simplification is a fundamental task in Natural Language Processing, aiming to replace complex words with simpler synonyms while preserving the original meaning of the text. This task is crucial for improving the accessibility of texts, particularly for users with reading difficulties, second language learners, and individuals with lower literacy levels. In this paper, we present MultiLS-IT, the first dataset specifically designed for automatic lexical simplification in Italian, as part of the larger multilingual Multi-LS dataset. We provide a detailed account of the data collection and annotation process, including complexity scores and synonym suggestions, along with a comprehensive statistical analysis of the dataset. With MultiLS-IT, we fill a significant gap in the field of Italian lexical simplification, offering a valuable resource for developing and evaluating automatic simplification models. Our analysis highlights the diversity of complexity levels in the dataset and discusses the moderate agreement among annotators, underscoring the subjective nature of lexical complexity assessment.

Keywords

lexical simplification, lexical complexity prediction, Italian dataset, human annotations

1. Introduction

Lexical simplification is a highly complex task within Natural Language Processing, encompassing broader automatic text simplification efforts [1]. It is defined as the task of replacing complex words with simpler synonyms that are more accessible to speakers, while preserving the original text's meaning [2]. A complex word is one that is difficult for some readers to decode due to various characteristics that hinder comprehension [3, 4].

This area of research is of significant interest both socially and in computational applications. Socially, automatic simplification can enhance text comprehension for individuals with reading difficulties [5, 6], second language learners [7], those with cognitive disabilities [8], or individuals with lower literacy levels [9]. In general, making texts accessible to everyone is a democratic act, as it ensures that information and knowledge are available to all members of society, regardless of their reading ability or educational background [10].

From a computational perspective, it proves valuable for complex tasks such as machine translation [11], information retrieval [12], and summarisation [13] in addition to being an integral part of generic text simplification [1]. The ability to simplify text effectively can improve the performance of these applications by making the input data more uniform and easier to process [2].

Lexical simplification encompasses various subtasks [14]. The two most important ones are:

1. the prediction of word complexity, which involves identifying the words that need to be simplified [15];
2. the replacement of complex words with simple synonyms [16].

Lexical complexity prediction (1) normally involves assigning a complexity value to a lexical item in context, ranging from 0 to 1, where 0 represents maximum simplicity and 1 denotes complexity [4]. This approach is a more advanced evolution of the traditional binary Complex Word Identification (CWI) [3], which classified words simply as complex or not complex. By moving towards a gradualism approach, lexical complexity prediction provides a finer-grained, continuous assessment of word difficulty, allowing for more tailored simplification efforts.

The replacement of complex words with simpler synonyms (2) comprises three subtasks: the generation of substitutes, the ranking based on complexity, and the selection of the most appropriate substitute [14]. This multi-step process ensures that the chosen synonym not only reduces complexity but also fits seamlessly into the original context.

One of the major challenges for such a user-dependent and therefore complex task is the lack of extensive annotated linguistic resources needed to train and evaluate automatic simplification models [2, 4]. Annotated datasets are crucial for developing and testing algorithms that can perform these tasks accurately.

In this context, we present MultiLS-IT, which is, to the best of our knowledge, the first dataset specifically designed for automatic lexical simplification in the Italian language. This resource is part of a larger multilingual

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

laura.occhipinti3@unibo.it (L. Occhipinti)

0009-0007-8799-4333 (L. Occhipinti)

dataset, Multi-LS (Multilingual Lexical Simplification) [17], created for a shared task at the BEA workshop [18][1].

The main contributions of this work are:

- A detailed description of the data collection and annotation process of the Italian sub-dataset;
- A descriptive analysis including statistics and visualizations providing an overview of the dataset's characteristics;
- The establishment of a reference point for future research in lexical simplification for Italian.

With this work, we aim to fill a significant gap in lexical simplification research for Italian and provide a solid foundation for future studies and more effective lexical simplification technologies.

2. Related works

Most datasets developed for lexical simplification have primarily focused on a few languages, with English being the most resourced language [18]. In recent years, however, there has been notable progress in creating resources for other languages, such as Spanish, Portuguese, and Japanese, which has facilitated advancements in lexical simplification tasks for these languages. Despite these efforts, specific datasets for the Italian language have been notably absent, hindering the development of comprehensive lexical simplification systems for Italian.

Many of these valuable datasets have been developed within the context of various shared tasks. The first one was proposed for SemEval 2012 [19]. It addressed English lexical simplification and provided a platform for evaluating systems that could rank substitution candidates by simplicity, using a dataset enriched with simplicity rankings from second language learners.

The CWI task at SemEval 2016 [20] focused on predicting which words in a sentence would be considered complex by non-native English speakers, creating a new dataset of 9,200 instances and attracting significant participation.

Expanding to multiple languages, the BEA 2018 CWI shared task [21] included English, German, and Spanish, and introduced a multilingual task with French, promoting the development of models capable of classifying word complexity across different languages.

The IberLEF 2020 forum [22] advanced Spanish lexical simplification by providing binary complexity judgments over educational texts, contributing to the available resources for Spanish.

The SemEval 2021 shared task on lexical complexity prediction [15] offered datasets for both single words and multi-word expressions in English, emphasizing continuous complexity judgments rather than binary classifications.

The SimpleText workshop at CLEF [23], initiated in 2021, aims to improve the accessibility of scientific information by providing benchmarks for text simplification, further expanding resources for this task.

The TSAR-2022 shared task [16] provided extensive annotations for lexical simplification in English, Spanish, and Portuguese, allowing participants to predict simple substitutions for complex words.

These datasets have catalyzed significant research and development in the field. For instance, the availability of such resources has enabled the implementation of full lexical simplification pipelines [24, 25, 26].

The majority of these datasets have typically concentrated on individual sub-tasks within the simplification pipeline, such as complex word identification (or lexical complexity prediction) or substitute generation. This division often limits the ability to comprehensively address the entire lexical simplification process.

In this context, Multi-MLSP represents a significant advancement [17]. It serves as a foundational resource for the entire simplification pipeline, annotated for both complexity values and potential substitutes. By providing a well-structured and annotated dataset, Multi-MLSP facilitates comprehensive research and development in lexical simplification, addressing both complexity prediction and the generation of simpler substitutes[2].

Despite these advancements, Italian has lagged behind due to the lack of dedicated resources.

2.1. Lexical Simplification Research in Italian

Numerous studies have explored automatic simplification for Italian [27], and several parallel corpora have been developed within these research projects [28, 29, 30, 31]. These corpora provide a valuable foundation for implementing automatic models for text simplification by presenting original texts aligned with their simplified versions. However, they primarily focus on syntactic simplification rather than lexical simplification, limiting their utility for tasks that require detailed lexical annotations.

We attempted to extract the lexical simplifications present in the available corpora using text comparison between simple and complex sentences with the `difflib` library. The lack of annotations made the recognition of substitutions complex and required significant manual effort. From the exploration of these substitutions, how-

[1] While some general information about the entire dataset has already been published in these papers [17, 18], the detailed process of constructing the Italian resource has not been thoroughly discussed until now.

[2] The resource, including the Italian part, is available for download from https://github.com/MLSP2024/MLSP_Data.

Target	Context	Complexity	Substitutions
popolareggiante	Lo stile è molto popolareggiante, a volte quasi con ostentazione (specialmente in alcune canzoni, che sembrano costituite da centoni di proverbi popolari), ma senza per questo risultare affettato.	0.3	comune, popolare, pop, basilare, casereccio, popolaresco, schietto, semplice
ostentazione	Lo stile è molto popolareggiante, a volte quasi con ostentazione (specialmente in alcune canzoni, che sembrano costituite da centoni di proverbi popolari), ma senza per questo risultare affettato.	0.12	esibizione, sfoggio, esagerazione, esibizionismo, sfacciataggine, presunzione
affettato	Lo stile è molto popolareggiante, a volte quasi con ostentazione (specialmente in alcune canzoni, che sembrano costituite da centoni di proverbi popolari), ma senza per questo risultare affettato.	0.52	costruito, forzato, ricercato, artefatto, artificioso, complesso, esagerato, falso, finto, innaturale, pomposo, preciso, pretenzioso, sdolcinato, studiato

Table 1
Examples of a MultiLS-IT sentences with target words and their substitutions.

ever, we realized that the steps of lexical simplification have never been truly systematized.

The only resource used to identify complex words and potential simpler substitutes has been *Nuovo Vocabolario di Base* [32], a dictionary of common Italian words. This resource, although fundamental and significant for the Italian language, is primarily built on the basis of word frequency. However, as we know from the literature [33], we cannot consider only a single measure, such as frequency, as a comprehensive parameter of complexity.

Furthermore, this resource, due to its nature as a static list, has inherent limitations in identifying complex words and generating suitable substitutes. For instance, consider the word *abolizione* (abolition), which is not included in De Mauro's basic vocabulary list, whereas its verb counterpart *abolire* (to abolish) is present. Speakers familiar with the meaning of *abolire* would likely comprehend *abolizione* relatively easily, deducing its meaning as the action or process of abolishing. This example underscores the limitation of solely relying on predefined reference lists, as speakers can understand logically connected words within their lexicon.

Given this scenario, there is a clear need for more comprehensive and annotated datasets that specifically address lexical simplification in Italian.

3. Dataset

MultiLS-IT is the Italian portion of a broader multilingual dataset, MultiLS. The overall dataset comprises 10 different languages: Catalan, English, Filipino, French, German, Italian, Japanese, Sinhala, Portuguese, and Spanish. To ensure consistency across the sub-datasets for each language, shared guidelines were established [17][3]. This section will outline the key aspects specific to the construction of the resource for Italian.

MultiLS-IT comprises 200 distinct contexts, each containing 3 target words. This design means that each sentence is repeated 3 times, as illustrated in Table 1, with each repetition focusing on a different target word. Consequently, the dataset includes a total of 600 sentences, corresponding to 600 target words.

For each target word, the dataset provides an average complexity value. This value is calculated by aggregating the complexity ratings assigned by individual annotators.

Additionally, the dataset includes a series of substitute words for each target word. These substitutes are ordered primarily by the frequency with which they were suggested by the annotators. In cases where multiple substitutes have the same frequency, they are listed alphabetically.

3.1. Data Preparation

For the construction of the MultiLS-IT dataset, we started by selecting the first 200 Italian words as outlined in the guidelines. The chosen words represent single lexical units, thus multi-word expressions were excluded[4].

The selection process ensured that the words were sufficiently complex to justify lexical complexity annotation and that simpler substitutes could be found within the context. Each target word required a minimum of 10 annotators.

Prior to selecting the words, we chose texts for the corpus. Given that the shared task, in the context of which this dataset was constructed, focused on educational applications, we selected texts related to educational settings, specifically Italian literature. This choice was reinforced by the importance of lexical simplification tasks in educational contexts, such as schools.

To ensure privacy and copyright compliance, texts from Wikimedia, specifically Wikibook and Wikiquote, were used. These texts are released under the Creative Commons Attribution-ShareAlike 3.0 license, allowing for use and sharing. We maintained a balanced ratio by selecting 50% of the texts from Wikibook and 50% from Wikiquote, as indicated in [18].

[3]The full guidelines are available at: https://github.com/MLSP2024/MLSP_Data/blob/main/MLSP%20Shared%20Task%20%40%20BEA%202024%20\protect\discretionary{\char\hyphenchar\font}{}{}%20Annotation%20Guidelines%20\protect\discretionary{\char\hyphenchar\font}{}{}%20V1.0.pdf.

[4]The guidelines provided two options for selecting words: we could either translate part of a sample list of 200 English words provided, or use this list as a guide to understand the type and distribution of words to select. We opted for the second approach, selecting the Italian words independently while using the English list only as a reference.

Web material extraction was carried out using BootCat [34], a tool that allows for automated collection of texts from the web.

To ensure the dataset reflected modern Italian usage, we applied specific filters to exclude archaic or outdated terms. We configured BootCat to focus on texts from the 20th century by using keywords such as '20th-century Italian literature', 'authors', 'female authors', and 'writers'. These filters helped us target contemporary Italian language and avoid the inclusion of words or expressions that are no longer in common usage. Through this approach, we ensured that the vocabulary extracted was relevant for current readers and aligned with modern Italian linguistic practices.

We employed a binary classifier developed for Italian CWI to select the words. The Random Forest model, detailed in [35], classifies words as simple (0) or complex (1) using various linguistic parameters to define lexical complexity.

The model was trained on a dataset comprising 13,319 words, labeled as simple or complex. To avoid subjective choices, this list of words was created based on linguistic resources related to L2 learning, ensuring an objective selection process. It is important to note that the complexity classification was done without considering the context in which the words appear due to the lack of available resources. This dataset includes features such as word frequency from two corpora (ItWac [36] and Subtlex-it [37]), word length, syllable count, vowel count, stop word identification, number of senses, POS tags, number of morphemes, morphological density, and the frequency of lexical morphemes. These metrics are commonly used because they have a significant impact on lexical complexity [38]. Additionally, pre-trained word embeddings from fastText were incorporated to enhance the model's predictions. The model underwent rigorous validation, demonstrating strong performance in accuracy, precision, recall, and F1 score. The classifier effectively utilized the combined linguistic features and word embeddings, providing a robust method for predicting word complexity.

This model was applied to the corpus of educational texts. To select the 200 words, we observed the complexity probabilities assigned by the model and chose those with the highest probabilities, ensuring that they allowed for easy identification of simpler synonyms.

For each sentence, in addition to the primary target word, we selected two additional content words to ensure a balanced representation of lexical complexity within the context. These words were chosen based on their semantic relevance to the sentence and their potential for simplification, meaning they could plausibly be replaced with simpler synonyms. The aim was to cover a range of complexity levels, avoiding an over-representation of either very simple or overly complex words.

The selection of the two additional words involved a manual search for content words—nouns, verbs, or adjectives—that could be substituted without altering the meaning or coherence of the sentence. In cases where multiple suitable content words were identified, we prioritized those for which a higher number of simpler substitutes could be found, applying the same approach used for the primary target word.

If a sentence did not allow for the selection of all three target words with suitable substitutions, it was excluded to ensure consistency across the dataset. This method guaranteed that all selected words were valid candidates for lexical simplification and provided a meaningful basis for analyzing word complexity and substitution potential.

3.2. Annotation

Our dataset provides a complexity rating for each target word, along with a set of synonyms perceived by annotators as simpler alternatives for replacement.

For the first task, annotators were instructed to assign a complexity rating based on 'how simple or complex the target word might be for a typical Italian native speaker'. Ratings were distributed on a 5-point Likert scale:

1. very easy - words that are very familiar
2. easy - words that are mostly familiar
3. neutral - when the word is neither difficult nor easy
4. difficult - words whose meanings are unclear but can be inferred from the context
5. very difficult - words that are very unclear.

The prediction of lexical complexity involves assigning a complexity score to a lexical item in context, typically ranging from 0 to 1. The aggregated complexity score, computed as the average of individual complexity ratings, initially ranged from 1 to 5 and was normalized using the min-max function following the Complex 2.0 format [39] as provided by the guidelines. The resulting scores were rounded to the nearest two decimal places.

For the second task, annotators were asked to suggest 1 to 3 synonyms that could replace the target word with simpler alternatives, aiming to enhance sentence comprehension. The substitutions were selected to ensure that the meaning of the original word and the overall context was preserved, and that the substitution was easier to understand than the original target. If the annotator could not find a simpler substitute, they were instructed to enter the target word itself as the suggestion to indicate that the term is the simplest word.

Specific instructions were provided to the annotators for the Italian dataset to avoid further complicating the already challenging task of finding suitable synonyms. It was permissible to disregard gender agreement within

the context. Additionally, pronominal verbs were to be treated as single entities that could be replaced by other types of verbs. For example, *mobilitarsi* (to mobilise oneself) could be substituted with *agire* (to act).

To ensure dataset robustness, a minimum of 10 annotations per word was required. Both complexity rating and synonym suggestion tasks were assigned to the same group of annotators for consistency.

Data collection was facilitated through Google Forms, where annotators evaluated sentences and proposed substitutions. We distributed 20 unique forms, each containing 30 sentences, and automated data compilation using Google App Script. Distribution channels included social media platforms like Instagram and Facebook, along with direct outreach to native speakers for participation.

Additionally, manual quality control was performed to ensure the reliability of the annotations. This included checking that annotators had used the full range of annotations and verifying that the complexity judgments were consistent with those of other annotators. For synonym suggestions, we checked the suitability of the substitutions within the context and monitored the frequency with which annotators were unable to find a simplification.

In total, 215 annotators participated, ensuring diverse and comprehensive representation. The metadata summarizing annotator demographics is presented in Table 2.

Age	36.39 (11.23)
Years in education	17.33 (3.27)
Nr. of L2-languages	2.17 (0.93)
Hours reading/week	7.39 (6.96)
Number of native annotators	215
L1-languages	Italian

Table 2
Average and standard deviation of Italian annotators' metadata.

This structured approach ensured data quality and reliability, crucial for subsequent analyses and computational model development in lexical complexity research.

3.3. Inter-Annotator Agreement

To evaluate the reliability of the complexity ratings, we calculated the inter-annotator agreement. This was done by assessing the consistency of the complexity scores assigned by different annotators to the same target words.

Given that our dataset consists of ordinal data representing complexity values ranging from 1 to 5, we employed Spearman's rank correlation coefficient to measure agreement. Spearman's correlation is appropriate for ordinal data as it assesses the strength and direction of the association between two ranked variables without assuming a linear relationship.

We calculated the Spearman correlation coefficient for each pair of annotators, using the `spearmanr` function from the `scipy.stats` module. This process was repeated for all possible annotator pairs within each of the 20 Google Forms, each annotated by at least 10 annotators. For each form, we then calculated the mean Spearman correlation coefficient to summarize the level of agreement among annotators for that form.

The overall mean of the Spearman correlation coefficients across all forms provides a single numerical measure of inter-annotator agreement for the entire dataset. This value is 0.4230.

The inter-annotator agreement value indicates a moderate level of consistency among annotators in their complexity ratings. This reflects the inherent subjectivity in assessing lexical complexity but also highlights the general alignment in annotators' judgments.

The process of finding and suggesting synonyms is inherently more variable and subjective, making it difficult to measure agreement in the same statistical manner as for ordinal complexity ratings.

3.4. Statistical Analysis

To gain a comprehensive statistical overview of our corpus, we calculated key metrics including the distribution of complexity values and the average length of sentences. This analysis provides insights into the characteristics of the dataset, which are essential for understanding the nature of the lexical simplification task.

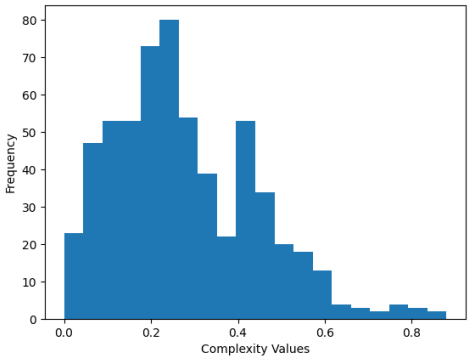

Figure 1: Distribution of complexity values.

The distribution of complexity values in the MultiLS-IT dataset is summarized as follows: the average complexity score across all target words is 0.276, with a standard deviation of 0.168. The range of complexity values spans

from 0.0 to 0.88. This distribution is visualized in Figure 1.

Additionally, we analyzed the sentence lengths within the dataset. The average sentence length is 29.30 words, with a standard deviation of 10.36 words. This measure helps in understanding the context provided for each target word, which is crucial for annotators when assigning complexity scores and suggesting simpler synonyms.

Furthermore, we investigated the correlation between sentence length and word complexity. The correlation coefficient between these two variables is 0.11, indicating a very weak relationship. This suggests that the complexity of a word is not significantly influenced by the length of the sentence in which it appears.

4. Conclusions

In this study, we present MultiLS-IT, the first dataset specifically designed for automatic lexical simplification in Italian. As part of the larger Multi-LS dataset, it addresses a significant gap in resources for lexical simplification in Italian. Despite its limited size, we believe that MultiLS-IT offers a valuable starting point for the development and evaluation of automatic simplification models. Our detailed description of the data collection and annotation process, including complexity ratings and synonym suggestions, provides a protocol that we hope will be followed and extended to increase the resources available for the Italian language.

Our analysis revealed that the average complexity score of all target words is 0.276, with a standard deviation of 0.168, highlighting the range of complexity levels within the dataset. Including more diverse and complex contexts would provide a richer resource for training and evaluating simplification models.

The inter-annotator agreement value of 0.4230 reflects a moderate level of consistency among annotators, emphasizing the inherent subjectivity in assessing lexical complexity. This relatively low value highlights the need to increase the sample size of both the dataset and the number of annotators to obtain more robust results.

Future work should focus on expanding the dataset to include a greater variety of texts and more annotators to improve the reliability and generalizability of the results. Our goal is to create broader resources that enable the development of robust and effective lexical simplification technologies that can improve text accessibility and comprehension for a wide range of readers.

In conclusion, while MultiLS-IT represents a significant step forward in the field of lexical simplification for Italian, there is still considerable potential for growth. Expanding the dataset to include a broader range of texts, increasing the number of annotators, and refining the annotation guidelines are all crucial steps toward improving the dataset's quality. Additionally, the application of more advanced computational models and the exploration of real-world use cases will further contribute to the development of sophisticated tools for lexical simplification. We hope that this dataset will serve as a foundation for future research and development in automatic simplification, ultimately making information more accessible and comprehensible to all.

References

[1] H. Saggion, G. Hirst, Automatic text simplification, volume 32, Springer, 2017.

[2] G. Paetzold, L. Specia, Lexical simplification with neural ranking, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, 2017, pp. 34–40. URL: https://aclanthology.org/E17-2006.

[3] M. Shardlow, 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, 2013, pp. 103–109.

[4] K. North, M. Zampieri, M. Shardlow, Lexical complexity prediction: An overview, ACM Computing Surveys 55 (2023) 1–42.

[5] D. De Hertog, A. Tack, Deep learning architecture for complex word identification, in: Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications, 2018, pp. 328–334.

[6] S. Stajner, Automatic text simplification for social good: Progress and challenges, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 2637–2652. URL: https://aclanthology.org/2021.findings-acl.233. doi:10.18653/v1/2021.findings-acl.233.

[7] J. S. Lee, C. Y. Yeung, Personalizing lexical simplification, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 224–232.

[8] X. Chen, D. Meurers, Linking text readability and learner proficiency using linguistic complexity feature vector distance, Computer Assisted Language Learning 32 (2019) 418–447.

[9] W. M. Watanabe, A. C. Junior, V. R. Uzêda, R. P. d. M. Fortes, T. A. S. Pardo, S. M. Aluísio, Facilita: reading assistance for low-literacy readers, in: Proceedings of the 27th ACM international conference on Design of communication, 2009, pp. 29–36.

[10] H. Saggion, J. O'Flaherty, T. Blanchet, S. Sharoff,

S. Sanfilippo, L. Muñoz, M. Gollegger, A. Rascón, J. L. Martí, S. Szasz, et al., Making democratic deliberation and participation more accessible: the idem project, in: SEPLN – CEDI 2024 Seminar of the Spanish Society for Natural Language Processing - 7th Spanish Conference on Informatics., 2024.

[11] S. Štajner, M. Popović, Can text simplification help machine translation?, in: Proceedings of the 19th Annual Conference of the European Association for Machine Translation, 2016, pp. 230–242.

[12] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, G. Hullender, Learning to rank using gradient descent, in: Proceedings of the 22nd international conference on Machine learning, 2005, pp. 89–96.

[13] Z. Cao, F. Wei, L. Dong, S. Li, M. Zhou, Ranking with recursive neural networks and its application to multi-document summarization, in: Proceedings of the AAAI conference on artificial intelligence, volume 29, 2015.

[14] M. Shardlow, Out in the open: Finding and categorising errors in the lexical simplification pipeline, in: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 1583–1590. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/479_Paper.pdf.

[15] M. Shardlow, R. Evans, G. H. Paetzold, M. Zampieri, SemEval-2021 task 1: Lexical complexity prediction, in: A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, X. Zhu (Eds.), Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, pp. 1–16. URL: https://aclanthology.org/2021.semeval-1.1. doi:10.18653/v1/2021.semeval-1.1.

[16] H. Saggion, S. Štajner, D. Ferrés, K. C. Sheang, M. Shardlow, K. North, M. Zampieri, Findings of the tsar-2022 shared task on multilingual lexical simplification, in: Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), 2022, pp. 271–283.

[17] M. Shardlow, F. Alva-Manchego, R. Batista-Navarro, S. Bott, S. Calderon Ramirez, R. Cardon, T. François, A. Hayakawa, A. Horbach, A. Hülsing, Y. Ide, J. M. Imperial, A. Nohejl, K. North, L. Occhipinti, N. Peréz Rojas, N. Raihan, T. Ranasinghe, M. Solis Salazar, M. Zampieri, H. Saggion, An extensible massively multilingual lexical simplification pipeline dataset using the MultiLS framework, in: R. Wilkens, R. Cardon, A. Todirascu, N. Gala (Eds.), Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 38–46. URL: https://aclanthology.org/2024.readi-1.4.

[18] M. Shardlow, F. Alva-Manchego, R. Batista-Navarro, S. Bott, S. Calderon Ramirez, R. Cardon, T. François, A. Hayakawa, A. Horbach, A. Hülsing, Y. Ide, J. M. Imperial, A. Nohejl, K. North, L. Occhipinti, N. P. Rojas, N. Raihan, T. Ranasinghe, M. S. Salazar, S. Štajner, M. Zampieri, H. Saggion, The BEA 2024 shared task on the multilingual lexical simplification pipeline, in: E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, Z. Yuan (Eds.), Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 571–589. URL: https://aclanthology.org/2024.bea-1.51.

[19] L. Specia, S. K. Jauhar, R. Mihalcea, Semeval-2012 task 1: English lexical simplification, in: * SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), 2012, pp. 347–355.

[20] G. Paetzold, L. Specia, SemEval 2016 task 11: Complex word identification, in: S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, T. Zesch (Eds.), Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 560–569. URL: https://aclanthology.org/S16-1085. doi:10.18653/v1/S16-1085.

[21] S. M. Yimam, C. Biemann, S. Malmasi, G. Paetzold, L. Specia, S. Štajner, A. Tack, M. Zampieri, 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, 2018, pp. 66–78.

[22] J. A. Ortiz-Zambranoa, A. Montejo-Ráezb, Overview of alexs 2020: First workshop on lexical analysis at sepln, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), volume 2664, 2020, pp. 1–6.

[23] L. Ermakova, P. Bellot, P. Braslavski, J. Kamps, J. Mothe, D. Nurbakova, I. Ovchinnikova, E. SanJuan, Overview of simpletext 2021-clef workshop on text simplification for scientific information access, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12, Springer, 2021, pp. 432–449.

[24] K. North, M. Zampieri, T. Ranasinghe, Alexsis-pt: A

new resource for portuguese lexical simplification, in: Proceedings-International Conference on Computational Linguistics, COLING, volume 29, 2022, pp. 6057–6062.

[25] L. Vásquez-Rodríguez, N. Nguyen, M. Shardlow, S. Ananiadou, Uom&mmu at tsar-2022 shared task: Prompt learning for lexical simplification, in: Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), 2022, pp. 218–224.

[26] K. North, T. Ranasinghe, M. Shardlow, M. Zampieri, Deep learning approaches to lexical simplification: A survey, arXiv preprint arXiv:2305.12000 (2023).

[27] D. Brunato, F. Dell'Orletta, G. Venturi, Linguistically-Based Comparison of Different Approaches to Building Corpora for Text Simplification: A Case Study on Italian, Frontiers in Psychology 13 (2022) 707630. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2022.707630/full. doi:10.3389/fpsyg.2022.707630.

[28] D. Brunato, F. Dell'Orletta, G. Venturi, S. Montemagni, Design and annotation of the first italian corpus for text simplification, in: Proceedings of The 9th Linguistic Annotation Workshop, 2015, pp. 31–41.

[29] S. Tonelli, A. Palmero Aprosio, F. Saltori, SIMPITIKI: a Simplification corpus for Italian, in: A. Corazza, S. Montemagni, G. Semeraro (Eds.), Proceedings of the Third Italian Conference on Computational Linguistics CLiC-it 2016, Accademia University Press, 2016, pp. 291–296. URL: http://books.openedition.org/aaccademia/1855. doi:10.4000/books.aaccademia.1855.

[30] D. Brunato, A. Cimino, F. Dell'Orletta, G. Venturi, Paccss-it: A parallel corpus of complex-simple sentences for automatic text simplification, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 351–361.

[31] M. Miliani, S. Auriemma, F. Alva-Manchego, A. Lenci, Neural readability pairwise ranking for sentences in Italian administrative language, in: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, 2022, pp. 849–866.

[32] T. De Mauro, I. Chiari, Il nuovo vocabolario di base della lingua italiana, Internazionale. [28/11/2020]. https://www. internazionale. it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana (2016).

[33] S. Bott, L. Rello, B. Drndarević, H. Saggion, Can spanish be simpler? lexsis: Lexical simplification for spanish, in: Proceedings of COLING 2012, 2012, pp. 357–374.

[34] M. Baroni, S. Bernardini, et al., Bootcat: Bootstrapping corpora and terms from the web, in: Proceedings of Fourth International Conference on Language Resources and Evaluation, LREC 2004, 2004, pp. 1313–1316.

[35] L. Occhipinti, Complex word identification for italian language: a dictionary-based approach, in: Proceedings of Clib24, Sixth International Conference on Computational Linguistics in Bulgaria, 2024, pp. 119–129.

[36] M. Baroni, S. Bernardini, A. Ferraresi, E. Zanchetta, The wacky wide web: a collection of very large linguistically processed web-crawled corpora, Language resources and evaluation 43 (2009) 209–226.

[37] D. Crepaldi, S. Amenta, M. Pawel, E. Keuleers, M. Brysbaert, Subtlex-it. subtitle-based word frequency estimates for italian, in: Proceedings of the Annual Meeting of the Italian Association For Experimental Psychology, 2015, pp. 10–12.

[38] K. Collins-Thompson, Computational assessment of text readability: A survey of current and future research, ITL-International Journal of Applied Linguistics 165 (2014) 97–135.

[39] M. Shardlow, R. Evans, M. Zampieri, Predicting lexical complexity in english texts: the complex 2.0 dataset, Language Resources and Evaluation 56 (2022) 1153–1194.