RALS: Resources and Baselines for Romanian Automatic Lexical Simplification

Fabian Anghel and **Petru Theodor Cristea** and **Claudiu Creangă** and **Sergiu Nisioi***
Human Language Technologies Research Center
Faculty of Mathematics and Computer Science
University of Bucharest
sergiu.nisioi@unibuc.ro

Abstract

We introduce the first dataset that jointly covers both lexical complexity prediction (LCP) annotations and lexical simplification (LS) for Romanian, along with a comparison of lexical simplification approaches. We propose a methodology for ordering simplification suggestions using a pairwise ranking approximation method, arranging candidates from simple to complex based on a separate set of human judgments. In addition, we provide human lexical complexity annotations for 3,921 word samples in context. Finally, we explore several novel pipelines for complexity prediction and simplification and present the first text simplification system for Romanian.[1]

1 Introduction

Text simplification is the process of transforming texts into variants that are simpler to understand by larger audiences or easier to process by existing NLP systems. Such initiatives promote literacy, facilitate effective communication, and enable equal access to information for individuals with diverse reading abilities or special needs (Zilio et al., 2020; Štajner, 2021; Gooding, 2022). These outcomes have broad-reaching advantages in sectors such as education, healthcare, legal documentation, government communication, online content, and beyond, ultimately enhancing social inclusivity and empowerment.

Unlike sentence simplification, which is typically modeled as a monolingual machine translation task (Specia, 2010; Nisioi et al., 2017; Dou et al., 2024), lexical simplification (LS) is specifically targeting particular lexical items to better control text generation and evaluation (Devlin, 1998; Carroll et al., 1998; De Belder and Moens, 2010; Glavaš and Štajner, 2015; Lee and Yeung, 2018; Sheang et al., 2022; Gooding and Tragut, 2022). In this way, the output can be guided towards specific target groups such as children or readers with different degrees of literacy.

End-to-end lexical simplification is typically divided into two equally challenging subtasks: **a) lexical complexity prediction (LCP)**, which assigns complexity scores to words (Yimam et al., 2018) and **b) lexical simplification**, which suggests simpler replacement words guided by LCP scores through candidate retrieval and re-ranking (North et al., 2023).

The series of workshops on lexical complexity prediction and lexical simplification (Specia et al., 2012; Paetzold and Specia, 2016; Yimam et al., 2018; Shardlow et al., 2021; Saggion et al., 2022; Shardlow et al., 2024b; Štajner et al., 2024) along with their shared tasks have nurtured a growing international interest in multilingual simplification. Several resources have been independently developed and used in these tasks that cover well-resourced languages such as Spanish (Ferrés and Saggion, 2022; Alarcon et al., 2023; Stajner et al., 2023), German (Ebling et al., 2022), Dutch (Hobo et al., 2023), Portuguese (Hartmann and Aluísio, 2020; North et al., 2022), Japanese (Kajiwara and Yamamoto, 2015; Kodaira et al., 2016; Ide et al., 2023), Chinese (Qiang et al., 2021b), and French (Billami et al., 2018; Pintard and François, 2020). Despite its potential broad impact, this task has received relatively little attention for medium and low-resource languages, where Italian, Catalan, Sinhala, or Filipino (Shardlow et al., 2024a), have remained relatively underexplored, and Eastern Romance languages such as Romanian are completely absent from landscape text simplification research. This scarcity of resources is not unique to Romanian (Codruț et al., 2024); many other languages worldwide face similar challenges due to their limited visibility on the global linguistic stage.

In this paper, we address the twin challenges of creating annotated datasets for lexical complexity

*Corresponding author.
[1] https://github.com/senisioi/RALS

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 31481–31492

prediction and lexical simplification in Romanian, a language currently lacking resources in this domain. Furthermore, we close this resource gap by constructing several end-to-end lexical simplification models specifically adapted to the particularities of Romanian.

2 Challenges and Related Work

Given the extremely low-resource setting, training end-to-end systems for Romanian comes with several challenges that cover both data construction and training techniques. To have comparable results across languages, LCP data should ideally be annotated with similar guidelines, for similar target groups, and comparable text genres across languages. One such attempt is the MultiLS dataset (Shardlow et al., 2024c) developed for the 2024 Shared Task (Shardlow et al., 2024b) which covers ten languages with similar methodological annotations, even though genres and annotator target groups differ across languages. Our data construction follows a multi-step approach involving human translation, machine translation, and multiple types of manual annotation, including self-assessment for lexical complexity prediction (LCP) and pairwise annotation for ranking simplification candidates.

The 2024 Shared Task (Shardlow et al., 2024b) showed that training systems from scratch on limited or synthetically generated data (Sastre et al., 2024) yields poorer results (Pearson $r \approx 0.4$) than zero-shot prompting with GPT-4 ($r \approx 0.6$) (Enomoto et al., 2024). However, the better-performing approach should not be considered a *silver bullet*, as prompting proprietary LLM systems to provide complexity score assessments carries practical risks, including privacy leaks, hallucinations, and uncontrolled output variability (Yao et al., 2024; Allen-Zhu, 2024).

An alternative approach that does not use LLMs was tested by Cristea and Nisioi (2024), who built cross-lingual LCP predictors using machine translations (MT) into English and back-translations; however, their experiments produced weak results ($r \approx 0.3$). MT has significant pitfalls: words that are easy in one language might not be easy in another, translations are rarely done word-by-word, and *translationese* is a language variety and lect with its own particularities (Blum and Levenston, 1978; Rabinovich et al., 2016). In our work, we incorporate both carefully curated translations and words sampled from original texts written in Romanian. We create a high-quality set of data. Each sentence and word is carefully chosen to have both original annotations and annotations comparable with those in English.

Regarding lexical simplification, rule-based methods consisting of different pipelines such as word-sense disambiguation, synonym reranking, and morphological operations (Paetzold and Specia, 2015; Ferrés et al., 2017) are less prevalent and are considered weaker than neural models because they depend on high-quality linguistic resources and robust pipelines. Nevertheless, the results reported by Saggion et al. (2022) at the TSAR Shared Task point out that some neural systems underperformed rule-based baseline.

In this work, we propose a hybrid solution that combines pre-existing synonym and morphological inflection dictionaries with a contextual embedding-based word-sense detector.

Neural network-based solutions dominate lexical simplification tasks, and we highlight two main types of systems (North et al., 2024): 1) masked or generative language models (Qiang et al., 2020, 2021a; Ferrés and Saggion, 2022; Sheang and Saggion, 2023), which perform word prediction and reranking, and 2) LLM-based instruction tuning (Baez and Saggion, 2023) or prompting of closed-source systems (Aumiller and Gertz, 2022; Enomoto et al., 2024). The latter has achieved the highest performance of any LS method across languages in both the 2022 and 2024 Shared Tasks (Saggion et al., 2022; Shardlow et al., 2024b).

Approaches based on masked language models for candidate suggestion (North et al., 2024) are generally ineffective for low-resource languages, as they often alter sentence meaning by suggesting simplifications from semantically related categories (e.g., *cat, dog, mouse; coffee, tea*). Furthermore, LLM prompting for candidate suggestion has difficulties in producing words in the correct inflected form and may lead to hallucinations (Cristea and Nisioi, 2024). Finally, proprietary systems come with cost restrictions, lack transparency, cannot be trusted with data requiring high privacy, and there is no guarantee of result consistency.

Our work addresses several of these challenges and offers a comparative analysis of lexical simplification methods.

	Sentence length				Complexity			
	En	**HT**	**WT**	**RoLCP**	**En**	**HT**	**WT**	**RoLCP**
mean	22.82	24.46	24.53	27.58	0.24	0.13	0.28	0.23
std	7.74	8.62	8.43	26.59	0.19	0.25	0.21	0.23
min	7	9	6	2	0.02	0	0	0
max	45	49	42	318	0.93	1	1	1
no. samples	569	569	1,765	1,587	569	569	1,765	1,587
no. sentences	190	190	751	274	190	190	751	274

Table 1: Statistics comparing the English and Romanian Human Translation (HT) sentences (569 samples), the Romanian Word-level translation (WT) datasets (1765 samples), and the new RoLCP data (1,587 samples). In total, 3,921 annotated samples for LCP on Romanian. The annotated complexity for words occurring in original texts resemble closer the distribution of the similar word annotations in original English.

3 Data Collection

3.1 Lexical Complexity Dataset

The English portion of the MultiLS dataset (Shardlow et al., 2024c) contains 569 word–sentence sample pairs sourced from Wikibooks. For each sentence, three words are annotated on a scale of 1 to 5, from simple to complex, through crowd-sourcing.

We construct three Romanian subsets:

1. HT: a direct counterpart, created by human translation (HT) of all 569 samples into Romanian and carefully aligning the most appropriate target word with the original English annotated equivalent;

2. WT: 1,765 samples identified in the Representative Corpus of Romanian (Midrigan Ciochina et al., 2020) using the same set of words as in the HT dataset, with the aim of testing whether sentences containing word translations (WT) offer better representativeness (see Appendix D);

3. RoLCP: 1,587 new samples containing words not included in HT or WT, selected based on frequency distributions, annotated in sentences drawn from diverse Romanian texts, including Wikipedia articles, popular science, literature, institutional documents, and argumentative essays.

For all three subsets the annotators are university students, similar to the annotations from MultiLS Shardlow et al. (2024c). We recruit a total of 90 native Romanian young adults with backgrounds in history, linguistics, and computer science. Using the Labelbox platform,[2] we present each sentence and target word in a randomized trial. Annotators assign one of five categorical labels: *very familiar*, *simple*, *neutral*, *difficult*, or *very difficult* to each target word. Before annotating the data, a trial of 15 samples is provided for practice and to explain the annotation guidelines. Following the Complex2.0 guidelines (Shardlow et al., 2022), we convert these labels to numerical values in the range [0, 1] and compute the average across annotators. Each word receives an average of 7.5 annotations, resulting in an overall inter-annotator agreement of Krippendorff's $\alpha = 0.37$ (with quadratic weights). Since lexical complexity is subjective and the group of annotators is heterogeneous, we do not expect complete agreement. Each participant rates word complexity based on personal judgment, and the final score for each word is the mean of all individual ratings. Table 1 and Figure 2 in the Appendix indicate a distribution shift in complexity following the translation of the English dataset into Romanian. Consistent with expectations and the findings from the MLSP dataset (Shardlow et al., 2024a), the English data exhibits a strong negative correlation (-0.74) between Zipf frequency (Speer, 2022) and complexity. This confirms the general principle that more frequent words tend to be simpler. By contrast, the Romanian datasets deviate from this pattern, containing a larger proportion of words annotated as simple, with a moderately negative correlation ($-0.53 \pm .01$) with Zipf frequency.

3.2 RoLS Dataset of Lexical Substitution Candidates

In the original English guidelines (Shardlow et al., 2024c), each annotator provides one to three replacement candidates for a given word. These candidates are then ranked by the number of suggestions, with the most frequently suggested word treated as the simplest. This design choice impacts evaluation metrics such as MAP@K, which are sensitive to the ordering of the top-K candidates. In addition, suggestion frequency may not always be a reliable indicator of simplicity. Annotator creativity, tied frequencies (observed in approximately

[2] https://labelbox.com

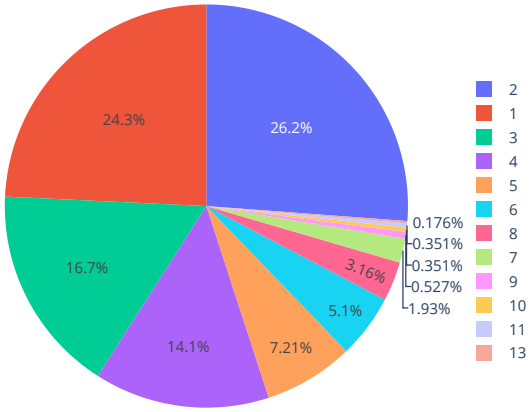

Figure 1: Percentage of examples with K lexical substitution candidates. Almost a quarter of sentences have no substitution candidates and almost half of sentences have at least 3 substitution candidates.

Train \Test	HT	WT	RoLCP
HT	$0.56 \pm .11$	0.54	0.52
WT	0.50	$0.68 \pm .04$	0.60
RoLCP	0.54	0.61	$0.71 \pm .05$

Table 2: Performance of models trained on one dataset and tested on the others. Diagonal entries report 10-fold average and std. grouped 10-fold cross-validation results for each dataset (HT, WT, RoLCP). The model shows limited generalization across datasets.

30% of the English data), and cases where less common words are actually simpler can all influence the results.

In our approach, two new annotators from the same target group suggest replacement candidates for the HT data without judging their simplicity. They can use external resources, including dictionaries and LLMs. The final list is the union of both sets of suggestions, verified by a third annotator. About 25% of target words have no suitable replacements, while nearly 50% have at least three (see Figure 1).

We then generate all possible sentence pairs from the suggested replacements and present them in random order to two additional annotators for pairwise simplicity judgments (Krippendorff's $\alpha = 0.53$). Each annotator reviews approximately 3,000 pairs. Using the method from Jerdee and Newman (2024), we apply a logistic Bradley–Terry model to compute sentence rankings based on pairwise comparisons (see Appendix B). This ranking methodology is more robust than the one used for English, as it requires annotators to read and compare two complete sentences with candidate words replaced in context.

4 Results

Lexical Complexity Prediction: we employ a simple Ridge regressor baseline with handcrafted features: 1. zipf_frequency from wordfreq library (Speer, 2022); 2. character length, number of vowels, approximate number of syllables from pyphen library;[3] 3. the number of immediate children in syntactic dependency parse and the static embeddings from spaCy `ro_core_news_lg` (Montani et al., 2023); 4. additional boolean features such as: is title, is entity, is sentence start, is sentence end regarded as markers of conceptual complexity (Stajner et al., 2020). We choose this approach for its simplicity and because it was one of the top-performing methods for Multilingual Lexical Complexity Prediction (Shardlow et al., 2024b).

We evaluate the models with the Pearson correlation coefficient, applying grouped 10-fold cross-validation so that no sentence in the training set appears in the test set, and we report cross-dataset evaluation scores in Table 2. Cross-validation shows that RoLCP is the most consistent dataset ($r = 0.71$), while HT is the most challenging ($r = 0.56$). Off-diagonal results indicate that models generalize moderately well across datasets, with WT $\rightarrow$ RoLCP (0.60) and RoLCP $\rightarrow$ WT (0.61) showing the best cross-dataset transfer. The results are comparable to those reported for other languages by Shardlow et al. (2024b): significantly lower than English (0.85), similar to Spanish (0.72) and German (0.71), close to French, Italian, and Catalan ($\simeq 0.62$), and higher than Filipino (0.56) and Sinhala (0.30). For all these languages, the reported scores rely on deep learning methods.

Lexical Simplification: the state-of-the-art approaches are strongly based on prompting external LLMs, as shown in the most recent Multilingual Lexical Simplification Pipeline (MLSP) Shared Task (Enomoto et al., 2024). We employ two open models: (1) Apertus-8B-Instruct-2509 from the Swiss AI Initiative (Hernández-Cano et al., 2025), a massively multilingual model in which Romanian is represented through the FineWeb corpus (Penedo et al., 2025) with 54 million tokens or 1.19% of its training data; and (2) RoLlama3.1-8B from OpenLLM-Ro (Masala et al., 2024), a model based on the Llama3.1-8B instruction tuned on Romanian data. The prompt (see Appendix E) is written in English, as it yielded better results than

[3] https://doc.courtbouillon.org/pyphen

Metric	Apertus-8B	RoLlama-8B	DexFlex *	GPT-4o	BERT-Ro
MAP@1	0.21	0.4	**0.41**	0.28	0.27
MAP@3	0.11	0.18	**0.31**	0.16	0.16
MAP@5	0.10	0.16	**0.3**	0.15	0.15
MAP@10	0.10	0.16	**0.33**	0.15	0.14
Potential@3	0.29	0.49	**0.6**	0.43	0.42
Potential@5	0.33	0.51	**0.67**	0.51	0.48
Potential@10	0.39	0.52	**0.7**	0.58	0.48
ACC@1@top_gold_1	0.11	**0.22**	0.15	0.16	0.14
ACC@2@top_gold_1	0.15	**0.26**	0.23	0.19	0.2
ACC@3@top_gold_1	0.18	0.27	**0.28**	0.23	0.26

Table 3: DexFlex consistently outperforms all other approaches across MAP and Potential metrics, showing strong robustness for synonym generation. While RoLlama achieves slightly better accuracy on top-gold metrics, it lags behind in overall ranking and coverage. The results suggest that a hybrid approach like DexFlex can be more effective than large general-purpose LLMs or fine-tuned BERT, especially when training on small datasets.

its Romanian counterpart, and includes the full sentence together with the target word. The models are tasked with generating a JSON object containing ordered candidate replacements. For comparison, we apply the same strategy using GPT-4o to evaluate the performance gap between open models and closed-source systems.

In addition, we train a Romanian BERT model (Dumitrescu et al., 2020) with cross-entropy loss to make the model predict each replacement candidate. The model is evaluated with 5-fold grouped cross-validation, and training is performed for 3 epochs with a batch size of 16 and a learning rate of 0.0006.

The DexFlex Framework is a quasi-rule-based simplification system developed as an extension of the spaCy library. It automates tasks such as grammatical processing and synonym suggestion, using information from the open-source dexonline dictionary.[4] DexFlex uses spaCy (Montani et al., 2023) to identify the part-of-speech and grammatical features (gender, number, and person) of the target word and selects synonyms from the dictionary based on the similarity of contextual word embeddings from BERT (Dumitrescu et al., 2020). The selected synonyms are properly inflected using dictionary information to be adequate in the context of the sentence (see Appendix A). The LCP pipeline is used exclusively in conjunction with DexFlex to rank the candidate synonyms.

Evaluation is carried out using three metrics (Shardlow et al., 2024b): Mean Average Precision (MAP@N) evaluates a model's precision by assessing how well it ranks the correct class among the top N predictions. Potential@N measures the likelihood of finding relevant items within the top N results, and Accuracy@N@top_gold_1 assesses how often the first most likely substitution appears within the top N highly probable predictions. Table 3 contains the evaluation results across these metrics. DexFlex consistently outperforms other approaches; however, the overall scores are considerably lower than what one might expect for a high-resource language (MAP@1 $\geq$ 0.72). Evaluation scores are in the ranges of other low resource languages (North et al., 2024) such as Sinhala ($\simeq$ 0.31) and Filipino ($\simeq$ 0.36).

5 Conclusions

Our work introduces the first text simplification resources for Romanian and highlights key challenges in developing tools for under-represented languages. Our analyses show that cross-lingual transfer of complexity scores is not a viable resource creation procedure, causing distribution shifts. Furthermore, a hybrid rule-based system with synonym and inflection dictionaries offers a state-of-the-art solution for Romanian lexical simplification. This method is both more ecologically sustainable and linguistically grounded, while also outperforming prompt-based approaches with the latest large language models. Finally, since lexical complexity can be reliably predicted using handcrafted features with performance comparable to LLM-based models (Shardlow et al., 2024b), we advocate for the development of simpler baseline models and for the integration of dictionaries into contemporary NLP pipelines wherever feasible.

[4] https://github.com/petruTH/DexFlex DEX is the acronym for the Explanatory Dictionary of Romanian, published and updated since 1975, available online in different variants at https://dexonline.ro.

6 Limitations

The creation of the datasets was a long-term process during which we developed the annotation standards, and, as such, the three Romanian LCP datasets have several differences: (1) the initial human translation and word translation datasets are completely annotated by five annotators, while the RoLCP dataset is annotated by a pool of 80 annotators, each contributing to random subsets of the data, resulting in 10 annotations per word. For the RoLS dataset, due to the time-consuming nature of identifying candidate substitutions and providing human judgments for the large number of pairwise candidate comparisons, this process was only completed for the HT dataset.

Since the target annotators are educated young adults aged 20-33, the complexity signals captured in the dataset may limit the generalizability of models trained on this data for broader real-world applications.

DexFlex has limitations, particularly in handling words with multiple parts of speech that share the same form. To address this, the framework uses spaCy to extract additional grammatical attributes in order for the correct part of speech and inflection to be applied to suggested synonyms. Certain nuances, such as distinguishing between similar articulate nouns like "copacul" (English: *the tree*) and "copacu" remain challenging due to database inconsistencies.

The total budget for running the experiments and conducting the data collection was 10$.

7 Ethics Statement

The manual labeling was carried out by volunteers who agreed to annotate the data at no cost, and we are grateful for their significant contribution. Participants were invited via email and some students used the collected data to develop their dissertations or to build in-class projects. The annotators agreed to publish labels along with the dataset under anonymity.

The texts we used for creating the dataset were sourced from platforms like Wikipedia, Wikibooks, and other public online sources. These sources either reside in the public domain or are published under permissive licenses (such as Creative Commons) or allow for academic fair use, i.e., small excerpts for research and the creation of derivative works.

We release our data and code under the CC BY-NC-SA 4.0 license.

Acknowledgments

We express our gratitude to the annotators whose labor was fundamental in building the Romanian dataset. We thank Oleksandra Kuvshynova, Mircea Marin, Radu Ciobanu, Anamaria Hodivoianu, and Ana Sabina Uban for their valuable support during the process of writing and refining this work. We would also like to thank the anonymous reviewers and the area chair for their constructive feedback.

This research is mainly supported by InstRead: Research Instruments for the Text Complexity, Simplification and Readability Assessment CNCS - UEFISCDI project number PN-IV-P2-2.1-TE-2023-2007 and by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 334906.

References

Rodrigo Alarcon, Lourdes Moreno, and Paloma Martínez. 2023. Easier corpus: A lexical simplification resource for people with cognitive impairments. *Plos one*, 18(4):e0283622.

Zeyuan Allen-Zhu. 2024. ICML 2024 Tutorial: Physics of Language Models. Project page: https://physics.allen-zhu.com/.

Dennis Aumiller and Michael Gertz. 2022. UniHD at TSAR-2022 shared task: Is compute all we need for lexical simplification? In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 251–258, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Anthony Baez and Horacio Saggion. 2023. LSLlama: Fine-tuned LLaMA for lexical simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 102–108, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Mokhtar Boumedyen Billami, Thomas François, and Núria Gala. 2018. Resyf: a french lexicon with ranked synonyms. In *27th International Conference on Computational Linguistics (COLING 2018)*.

Shoshana Blum and Eddie A Levenston. 1978. Universals of lexical simplification. *Language learning*, 28(2):399–415.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 workshop on integrating*

artificial intelligence and assistive technology, pages 7–10. Madison, WI.

Rotaru Codruț, Nicolae Ristea, and Radu Ionescu. 2024. Rodia: A new dataset for romanian dialect identification from speech. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 279–286.

E. Coleman et al. 2021. Preparing accessible and understandable clinical research participant information leaflets and consent forms: a set of guidelines from an expert consensus conference. *Research Involvement and Engagement*, 7(1):31.

Petru Cristea and Sergiu Nisioi. 2024. Archaeology at mlsp 2024: Machine translation for lexical complexity prediction and lexical simplification. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 610–617, Mexico City, Mexico. Association for Computational Linguistics.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Proceedings of the SIGIR workshop on accessible search systems*, pages 19–26.

Siobhan Devlin. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic databases*.

Yao Dou, Philippe Laban, Claire Gardent, and Wei Xu. 2024. Automatic and human-ai interactive text generation (with a focus on text simplification and revision). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 5: Tutorial Abstracts)*, pages 3–4.

Stefan Dumitrescu, Andrei-Marius Avram, and Sampo Pyysalo. 2020. The birth of romanian bert. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4328.

Sarah Ebling, Alessia Battisti, Marek Kostrzewa, Dominik Pfütze, Annette Rios, Andreas Säuberli, and Nicolas Spring. 2022. Automatic text simplification for german. *Frontiers in Communication*, 7:706718.

Taisei Enomoto, Hwichan Kim, Tosho Hirasawa, Yoshinari Nagai, Ayako Sato, Kyotaro Nakajima, and Mamoru Komachi. 2024. TMU-HIT at MLSP 2024: How well can GPT-4 tackle multilingual lexical simplification? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 590–598, Mexico City, Mexico. Association for Computational Linguistics.

Daniel Ferrés and Horacio Saggion. 2022. ALEXSIS: A dataset for lexical simplification in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3582–3594, Marseille, France. European Language Resources Association.

Daniel Ferrés, Horacio Saggion, and Xavier Gómez Guinovart. 2017. An adaptable lexical simplification architecture for major ibero-romance languages. *EMNLP 2017*, page 40.

Organisation for Economic Co-Operation and Development. 2013. *OECD skills outlook 2013: first results from the survey of adult skills*. Organization for Economic Co-operation and Development (OECD), Paris Cedex, France.

Organisation for Economic Co-operation and Development (OECD). 2020. *Improving educational equity in Romania*.

Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68, Beijing, China. Association for Computational Linguistics.

Sian Gooding. 2022. On the ethical considerations of text simplification. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 50–57, Dublin, Ireland. 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.

Nathan Siegle Hartmann and Sandra Maria Aluísio. 2020. Adaptação lexical automática em textos informativos do português brasileiro para o ensino fundamental. *Linguamática*, 12(2):3–27.

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, Ines Altemir Marinas, Mohammad Hossein Amani, Matin Ansaripour, Ilia Badanin, Harold Benoit, Emanuela Boros, Nicholas Browning, Fabian Bösch, Maximilian Böther, Niklas Canova, Camille Challier, Clement Charmillot, Jonathan Coles, Jan Deriu, Arnout Devos, Lukas Drescher, Daniil Dzenhaliou, Maud Ehrmann, Dongyang Fan, Simin Fan, Silin Gao, Miguel Gila, María Grandury, Diba Hashemi, Alexander Hoyle, Jiaming Jiang, Mark Klein, Andrei Kucharavy, Anastasiia Kucherenko, Frederike Lübeck, Roman Machacek, Theofilos Manitaras, Andreas Marfurt, Kyle Matoba, Simon Matrenok, Henrique Mendonça, Fawzi Roberto Mohamed, Syrielle Montariol, Luca Mouchel, Sven Najem-Meyer, Jingwei Ni, Gennaro Oliva, Matteo Pagliardini, Elia Palme, Andrei Panferov, Léo

Paoletti, Marco Passerini, Ivan Pavlov, Auguste Poiroux, Kaustubh Ponkshe, Nathan Ranchin, Javi Rando, Mathieu Sauser, Jakhongir Saydaliev, Muhammad Ali Sayfiddinov, Marian Schneider, Stefano Schuppli, Marco Scialanga, Andrei Semenov, Kumar Shridhar, Raghav Singhal, Anna Sotnikova, Alexander Sternfeld, Ayush Kumar Tarun, Paul Teiletche, Jannis Vamvas, Xiaozhe Yao, Hao Zhao Alexander Ilic, Ana Klimovic, Andreas Krause, Caglar Gulcehre, David Rosenthal, Elliott Ash, Florian Tramèr, Joost VandeVondele, Livio Veraldi, Martin Rajman, Thomas Schulthess, Torsten Hoefler, Antoine Bosselut, Martin Jaggi, and Imanol Schlag. 2025. Apertus: Democratizing open and compliant llms for global language environments. *Preprint*, arXiv:2509.14233.

Eliza Hobo, Charlotte Pouw, and Lisa Beinborn. 2023. "geen makkie": Interpretable classification and simplification of Dutch text complexity. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 503–517, Toronto, Canada. 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.

M. Nisbeth Jensen and A. M. Fage-Butler. 2016. Antenatal group consultations: Facilitating patient-patient education. *Patient Education and Counseling*, 99(12):1999–2004.

M. Jerdee and M. E. J. Newman. 2024. Luck, skill, and depth of competition in games and social hierarchies. *Science Advances*, 10(45):eadn2654. Epub 2024 Nov 6. PMID: 39504380; PMCID: PMC11540035.

Tomoyuki Kajiwara and Kazuhide Yamamoto. 2015. Evaluation dataset and system for japanese lexical simplification. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*, pages 35–40.

M. Kim, D. Suh, J. A. Barone, S.-Y. Jung, W. Wu, and D.-C. Suh. 2022. Health literacy level and comprehension of prescription and nonprescription drug information. *International Journal of Environmental Research and Public Health*, 19(11).

Tomonori Kodaira, Tomoyuki Kajiwara, and Mamoru Komachi. 2016. Controlled and balanced dataset for Japanese lexical simplification. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 1–7.

John Lee and Chak Yan Yeung. 2018. Personalizing lexical simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 224–232, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mihai Masala, Denis Ilie-Ablachim, Alexandru Dima, Dragos Georgian Corlatescu, Miruna-Andreea Zavelca, Ovio Olaru, Simina-Maria Terian, Andrei Terian, Marius Leordeanu, Horia Velicu, et al. 2024. "vorbești românește?" a recipe to train powerful romanian llms with english instructions. pages 11632–11647.

Ludmila Midrigan Ciochina, Victoria Boyd, Lucila Sanchez-Ortega, Diana Malancea_Malac, Doina Midrigan, and David P. Corina. 2020. Resources in underrepresented languages: Building a representative Romanian corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3291–3296, Marseille, France. European Language Resources Association.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. explosion/spaCy: v3.7.2: Fixes for APIs and requirements.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2024. Deep learning approaches to lexical simplification: A survey. *Journal of Intelligent Information Systems*, pages 1–24.

Kai North, Marcos Zampieri, and Tharindu Ranasinghe. 2022. Alexsis-pt: A new resource for portuguese lexical simplification. In *Proceedings-International Conference on Computational Linguistics, COLING*, volume 29, pages 6057–6062.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Computing Surveys*, 55(9):1–42.

Gustavo Paetzold and Lucia Specia. 2015. LEXenstein: A framework for lexical simplification. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 85–90, Beijing, China. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

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.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. Fineweb2: One pipeline to scale them all–adapting pre-training data processing to every language. *COLM*.

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.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, Yang Shi, and Xindong Wu. 2021a. Lsbert: Lexical simplification based on bert. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3064–3076.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. Lexical simplification with pretrained encoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8649–8656.

Jipeng Qiang, Xinyu Lu, Yun Li, Yunhao Yuan, and Xindong Wu. 2021b. Chinese lexical simplification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1819–1828.

Ella Rabinovich, Sergiu Nisioi, Noam Ordan, and Shuly Wintner. 2016. On the similarities between native, non-native and translated texts. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1881.

Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2022. Findings of the TSAR-2022 shared task on multilingual lexical simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 271–283, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Ignacio Sastre, Leandro Alfonso, Facundo Fleitas, Federico Gil, Andrés Lucas, Tomás Spoturno, Santiago Góngora, Aiala Rosá, and Luis Chiruzzo. 2024. RETUYT-INCO at MLSP 2024: Experiments on language simplification using embeddings, classifiers and large language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 618–626, Mexico City, Mexico. Association for Computational Linguistics.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Marcos Zampieri, and Horacio Saggion. 2024a. An extensible massively multilingual lexical simplification pipeline dataset using the MultiLS framework. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024*, pages 38–46, Torino, Italia. ELRA and ICCL.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Huelsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024b. The BEA 2024 Shared Task on the Multilingual Lexical Simplification Pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Huelsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Marcos Zampieri, and Horacio Saggion. 2024c. An Extensible Massively Multilingual Lexical Simplification Pipeline Dataset using the MultiLS Framework. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI)*.

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.

Kim Cheng Sheang, Daniel Ferrés, and Horacio Saggion. 2022. Controllable lexical simplification for English. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 199–206, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Kim Cheng Sheang and Horacio Saggion. 2023. Multilingual controllable transformer-based lexical simplification. *Procesamiento del Lenguaje Natural*, 71:109.

Lucia Specia. 2010. Translating from complex to simplified sentences. In *Computational Processing of the Portuguese Language: 9th International Conference, PROPOR 2010, Porto Alegre, RS, Brazil, April 27-30, 2010. Proceedings 9*, pages 30–39. Springer.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. 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)*, pages 347–355.

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Sanja Štajner. 2021. Automatic text simplification for social good: Progress and challenges. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652.

Sanja Stajner, Daniel Ibanez, and Horacio Saggion. 2023. LeSS: A computationally-light lexical simplifier for Spanish. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1132–1142, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Sanja Stajner, Sergiu Nisioi, and Ioana Hulpuș. 2020. CoCo: A tool for automatically assessing conceptual complexity of texts. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7179–7186, Marseille, France. European Language Resources Association.

Sanja Štajner, Horacio Saggio, Matthew Shardlow, and Fernando Alva-Manchego, editors. 2024. *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability*. EMNLP 2024, Miami, Florida.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211.

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.

Leonardo Zilio, Liana Braga Paraguassu, Luis Antonio Leiva Hercules, Gabriel Ponomarenko, Laura Berwanger, and Maria José Bocorny Finatto. 2020. A lexical simplification tool for promoting health literacy. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 70–76, Marseille, France. European Language Resources Association.

Sanja Štajner, Daniel Ferrés, Matthew Shardlow, Kat North, Marcos Zampieri, and Horacio Saggion. 2022. Lexical simplification benchmarks for english, portuguese, and spanish. *Frontiers in Artificial Intelligence*, 5:991242.

A DexFlex Pipeline

The pipeline used for this study contains several stages which are detailed below.

The first step in suggesting synonyms using the DexFlex framework is identifying the target word along with a series of additional information about it. More specifically, we use spaCy to find the current part of speech of the word and to get the necessary details regarding various grammatical attributes such as number, person, or gender.

The second step in the pipeline involves selecting synonyms from the dexonline database. For better accuracy of this process, we first establish the contextual meaning of the word by comparing the current sentence context with representative contextual examples of the alternatives found in the dexonline database. The contextual examples found in the dexonline database are stored as cached embeddings and aproximate nearest neighbour search is used to identify the meaning with the highest cosine similarity. The synonyms are retrieved as lemmas. Summary evaluations showed that this approach is reliable in correctly disambiguating words that have multiple meanings for Romanian, but we did not run exhaustive word-sense-disambiguation evaluation.

The third step in the pipeline involves bringing the substitution candidates to the correct inflected form. In this regard, we use the grammatical knowledge derived from spaCy `ro_core_news_lg` together with simple pre-defined rules to retrieve inflected forms from the dexonline database.

B Simplicity Ranking from Pair-wise Assessments

We use the methodology proposed by Jerdee and Newman (2024) to estimate a ranking from pairwise binary simplicity scores assigned by annotators to sentences.

Considering a set of n replacement candidates labeled by $i = 1 \cdot n$, assign to each a real score parameter $s_i \in [-\infty, \infty]$. Then the probability that i is simpler than j is assumed to be some function of the difference of their scores: $p_{ij} = f(s_i - s_j)$. The function $f(s)$ satisfies the following axioms: it is increasing in s, it tends to 1 as $s \to \infty$ and to 0 as $s \to -\infty$, and it is asymmetric about its mid-point at $s = 0$ with the form $f(-s) = 1 - f(s)$. The logistic function is a popular choice, which gives $f(s_i - s_j) = \frac{e^{s_i}}{e^{s_i} + e^{s_j}}$ also known as the Bradley-Terry model.

Now, suppose we observe m matches between n players. The outcomes of the matches can be represented by an $n \times n$ matrix A with element A_{ij} equal to the number of times player i beat player j. The probability of the complete set of ob-

served outcomes is $P(A|s) = \prod_{ij} f(s_i - s_j)^{A_{ij}} = \prod_{ij} \left(\frac{e^{s_i}}{e^{s_i} + e^{s_j}} \right)^{A_{ij}}$, where s is the vector with elements s_i.

We calculate a maximum a posteriori (MAP) estimate of the values of the scores as: $\hat{s} = argmax_s P(s|A) = argmax_s P(A|s)P(s)$, given a prior with the variance chosen as $\frac{1}{2}$: $P(s) = \prod_{i=1}^{n} \frac{1}{\sqrt{\pi}} e^{-s_i^2}$. According to Jerdee and Newman (2024), the MAP estimate always exists regardless of whether the interaction network is strongly connected or not and using a prior eliminates the need for normalization.

C Literacy in Romania

According to the adult literacy report conducted in 24 highly-developed countries (for Economic Co-Operation and Development, 2013; Štajner et al., 2022), 16.7% of the population, on average, cannot understand texts that go beyond a basic vocabulary. Furthermore, functional literacy in Romania remains among the lowest in the European Union, with around 40% of students not achieving baseline proficiency in at least one subject in PISA 2015 (for Economic Co-operation and , OECD).

Further research has correlated low literacy with health risks - limited understanding of medication instructions (Coleman et al., 2021), misinterpretation of drug treatment information (Kim et al., 2022; Jensen and Fage-Butler, 2016), and the inability to make informed decisions in following a treatment or reading medical information. Creating text simplification systems can have several long-term societal benefits. The lack of pre-existing research and annotated data for Romanian text simplification / complexity represents an opportunity to bring novel contributions and create guidelines for developing similar approaches to new languages.

D Romanian Original Datasets: WT and RoLCP

We use the Representative Corpus of Romanian (Midrigan Ciochina et al., 2020) consisting of a diverse set (21 different genres) of written texts and speech transcripts from Romania and Moldova. The entire corpus is split into sentences using the large Romanian spaCy (Montani et al., 2023) model.

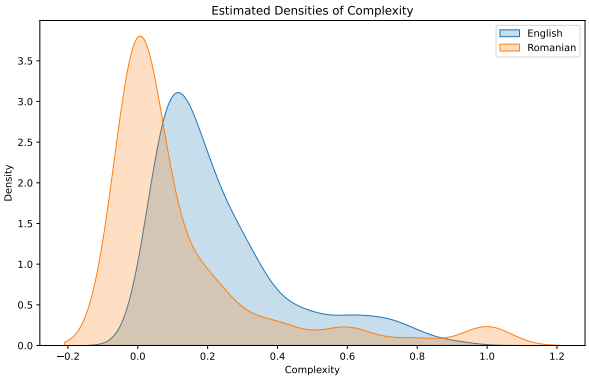

Figure 2: Kernel density estimations of the English vs. Romanian average complexity annotations. The translation process introduces shifts in the lexical complexity. A word that is considered medium complex in English has a Romanian translation that is perceived as simpler by native Romanian speakers.

Sentences containing each word are filtered to match the average length of our dataset. Because a word may have different meanings and functions depending on context, we apply the following procedure to construct the final sentence list. For each target word, we extract contextualized word embeddings using a Romanian BERT model (Dumitrescu et al., 2020) and project them into two dimensions with t-SNE. The resulting representations are clustered with KMeans, and the optimal number of clusters is selected according to the silhouette score. From each cluster, we sample 15% of the sentences: half are drawn from those closest to the centroid, representing prototypical usages, and half from those farthest away, capturing peripheral or atypical contexts. This yields a balanced subset that reflects diverse instances within each cluster of meaning (an example is shown in Figure 3). Finally, we manually review the selected samples, remove noisy sentences, and submit the remainder for explicit complexity annotation. The final dataset contains 1,765 sentences, with statistics summarized in Table 1. Its mean complexity is slightly higher than that of the English and human-translated datasets, and the difference is statistically significant ($p < 0.001$) according to a bootstrapping permutation test.

We construct a third set of annotations (dataset RoLCP) using sentences sourced from original Romanian texts, without constraints imposed by a predefined list of target words. To this end, we select 12 texts spanning diverse genres, including Wikipedia articles, popular science, literature, institutional documents, and argumentative essays.

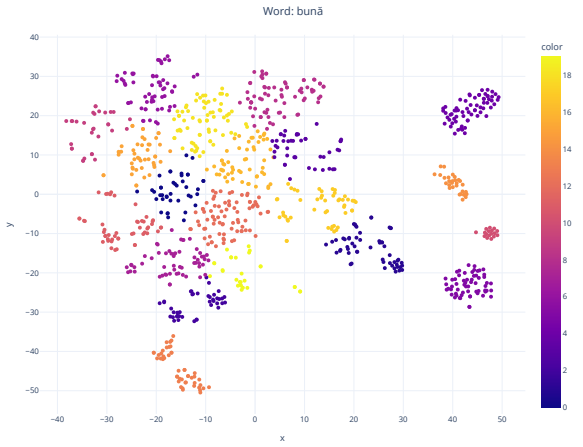

Figure 3: The different clusters for word "bună [en. *good*]" containing different meanings and context for the usage of the word. The clusters contain sentences with different collocations of the word (approximate translations: *good day*, *good decision*, *good food*, *good to go*, *good side of things*, *good will*, etc.). The sentence selection process incorporates both samples close to the centroid, representing prototypical usages, and outlier samples, reflecting less typical contexts.

Each text is sentence-split, and up to five target words per sentence are selected for annotation. The selection of target words is based on precomputed frequency statistics from a large Romanian corpus derived from (Speer, 2022), so that annotators assess both high-frequency and low-frequency words.

The pool of annotators have been randomly assigned different samples, yielding a total of 10 annotations per sample. The annotation process takes place in a lab in complete silence, annotators have been given a practice dataset of 15 sentences before beginning the actual process. During the lab-centered annotation process, annotators may ask questions and clarify corner cases with the supervisors.

E Lexical Simplification Prompt

```
Provide a list of 10 alternative simpler words (as a json object) that a child would understand easily to replace the word "ORIGINAL_WORD" in the context of the following sentence. It is mandatory to use suitable meanings for the context of the sentence and for the pattern of the answer to be displayed as a JSON with words as keys and complexity scores as values with all the 10 alternatives. Provide only words in "LANGUAGE". Sentence: "ORIGINAL".
```