Complexity and Indecision: A Proof-of-Concept Exploration of Lexical Complexity and Lexical Semantic Change

David Alfter
Gothenburg Research Infrastructure in Digital Humanities (GRIDH)
Department of Literature, History of Ideas and Religion
University of Gothenburg, Sweden
david.alfter@gu.se

Abstract

This paper explores the intersection of lexical complexity prediction and lexical semantic change detection. We investigate the potential connection between changes in lexical complexity and lexical semantics, aiming to uncover how these two aspects of language evolution are intertwined. Our findings indicate that lexical complexity models human annotator uncertainty surprisingly well. Further, we find a moderate correlation between changes in lexical complexity and graded lexical semantic change. This highlights the potential for leveraging lexical complexity for lexical semantic change detection.

1 Introduction

Though seemingly distinct, the fields of lexical complexity prediction and lexical semantic change detection share surprising points of contact. One predicts the inherent difficulty of words (Lexical Complexity Prediction, LCP; North et al. 2023), while the other tracks shifts in meaning and usage (Lexical Semantic Change Detection, LSCD; Tahmasebi et al. 2021). Despite starting with words as their foundational unit, both have gravitated towards considering individual word *senses* thanks to advancements in transformer models (Vaswani et al. 2017) and contextualized word embeddings. While LSCD inherently deals with this concept, research in LCP suggests different senses within a word exhibit varying complexities (Crossley et al. 2010; Alfter 2021; Shardlow et al. 2022). It has also been noted that the manual annotation of both LSCD data (e.g., whether a word has the same, closely/distantly related, or unrelated sense in two given sentences) and LCP data (how complex a word is in a given sentence) is quite subjective (Shardlow et al., 2021; Schlechtweg et al., 2021). Given the shared focus on contextual meaning and the inherent subjectivity of the tasks, we postulate a potential link.

In this paper, we specifically explore whether lexical complexity can explain human uncertainty in annotation. Utilizing human judgments on semantic closeness of words in sentences, we analyze if lexical complexity differences between sentences correlate with annotator indecision. As a downstream task, we also look at whether lexical complexity can directly predict lexical semantic change.

The rest of the paper is structured as follows: in section 2 we contextualize our work and highlight the commonalities and gap in communication between these disciplines. In section 3, we detail the methodological framework, including the dataset and experimental design. In section 4 we present the key findings of our experiments. In section 5, we interpret our results in a broader context, discussing their implications and potential future directions.

2 Related Work

2.1 Lexical complexity prediction

Lexical complexity prediction tries to identify the *complexity* of words in a text, with downstream tasks such as text simplification of various genres (e.g., medical texts (Deléger and Zweigenbaum, 2009), legal texts (LoPucki, 2014)) for various groups (e.g., children (De Belder and Moens, 2010), language learners (Petersen and Ostendorf, 2007), people with disabilities (Devlin, 1998; Chung et al., 2013)). Lexical complexity prediction has been explored in several shared tasks and several languages: the Complex Word Identification 2016 shared task for English (Paetzold and Specia, 2016a), the Complex Word Identification 2018 shared task for English, Spanish, German and French(Yimam et al., 2018), the ALexS 2020 shared task for Spanish (Ortiz-Zambranoa and Montejo-Ráezb, 2020), and the Lexical Complexity Prediction 2021 shared task for English

Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change, pages 137–143

(Shardlow et al., 2021).

Early tasks focused on binary complexity ("is the word complex or not?") while later tasks focus on graded complexity ("how complex is the word?"). While early system relied on feature engineering (e.g., Paetzold and Specia 2016b; Gooding and Kochmar 2018), later approaches use transformer-based models (e.g., Pan et al. 2021; Yaseen et al. 2021) or a combination of classical features and transformers (Paetzold, 2021). However, fully feature engineered systems still perform almost on-par with transformer-based systems; the best fully feature engineered system scored third place in the task (Shardlow et al., 2021; Agarwal and Chatterjee, 2021; Mosquera, 2021).

2.2 Lexical semantic change detection

Lexical semantic change detection tries to identify words that have undergone shifts in meaning over time, mainly as a task in itself, but also for downstream tasks such as OCR error correction (Morsy and Karypis, 2016) or document similarity computation (Chiron et al., 2017).

Lexical semantic change detection is an unsupervised tasks, and systems to detect this change generally use techniques such as word2vec (Mikolov and Dean, 2013) to represent words in a continuous vector space, allowing for the analysis of semantic similarities and changes over time (Tahmasebi et al., 2021). Other systems use co-occurrence information to build matrices and measure similarity between words based on their contexts (Sagi et al., 2009). Pointwise mutual information scores and cosine similarity are often employed to track changes in co-occurrence patterns over time to uncover how word meanings evolve and shift across different contexts (Teh et al., 2004; Gulordava and Baroni, 2011). Some methods use topic modeling to partition information based on word senses, allowing for the detection of sense changes over time (Lau et al., 2012). Topics are interpreted as senses, and new induction methods aim to infer sense and topic information jointly (Wang et al., 2015). Techniques such as word sense induction or discrimination aim to identify different senses of a word and track changes in these senses over time. Recent works use transformer-based models and average pairwise distance and prototype distance to detect change (Cassotti et al., 2023).

Lexical semantic change detection has been explored in several shared tasks for various languages: the SemEval 2020 task 1 on unsupervised lexical semantic change detection for English, German, Swedish and Latin (Schlechtweg et al., 2020), DIACR-Ita for Italian (Basile et al., 2020), RuShiftEval for Russian (Kutuzov and Pivovarova, 2021), and the SemEval 2022 task on semantic change discovery and detection in Spanish (Zamora-Reina et al., 2022).

The main common point between lexical complexity prediction and lexical semantic change lies in polysemy. Polysemy is a strong predictor of lexical complexity (Gala et al., 2013; Alfter and Volodina, 2018), as more polysemous words can occur in more varied contexts. As the context influences the specific meaning of the word, we expect the complexity to vary more strongly if the possible contexts are more numerous. In unsupervised lexical semantic change detection, the context is crucial in determining whether a word occurs in a given sense, and the degree of polysemy (and its change) is directly linked to lexical semantic change. To the best of our knowledge, there is no prior work investigating the role of lexical complexity in lexical semantic change.

3 Methodology

3.1 Data

For lexical semantic change detection, we use the English data from the Diachronic Word Usage Graph (DWUG) dataset (Schlechtweg et al., 2021) used in the SemEval 2020 shared task on unsupervised lexical semantic change detection (Schlechtweg et al., 2020). The shared task covered English, German, Swedish, and Latin, with labels for graded lexical semantic change as well as binary change. The English portion of the data set covers 46 target words, each with 200 sentences split across two time spans (100 per time span). The data was manually annotated using a Word-in-Context approach where annotators are asked to rate the semantic closeness of a word in two sentences on a scale from 1 (unrelated) to 4 (identical); as an additional annotation option, there is 0 which means 'cannot decide'. These judgments are then clustered to derive sense clusters, based on which a graded lexical semantic change score $\in [0-1]$ is computed (Schlechtweg et al., 2020).

For lexical complexity prediction, we use the data from the SemEval 2021 shared task on lexical complexity prediction (Shardlow et al., 2021). This data is only available for English. It contains about 9000 words from three genres (biblic, parliamen-

tary, medical). The data was manually annotated using a five-point Likert scale from 1 (very easy) to 5 (very difficult). These judgments were aggregated and normalized into the range $[0-1]$.

3.2 Models

We first fine-tune a model for lexical complexity on the provided training data from the 2021 shared task, evaluating on the trial data and testing on the test data. We take inspiration from Pan et al. (2021), the top performing team at the shared task, and prepend the word to the sentence (see Figure 1), but we omit the genre information, since this information is not available for the semantic change detection data.

Pan et al.: [CLS] genre word [SEP] sentence
Our input: [CLS] word [SEP] sentence

Figure 1: Illustration of Pan et al. (2021)'s input format versus our input format

As a proof of concept experiment, we do not follow Pan et al. (2021) and other teams in creating an ensemble of transformers for prediction; instead, we use a single RoBERTa-base model.[1] We train the model for 20 epochs with the R^2 objective and early stopping.

We then apply the fine-tuned model to the lexical semantic change data set: for each target word, we predict the complexity for each context it occurs in. We then calculate the average complexity for time span 1 (C^{t1}_{avg}) and time span 2 (C^{t2}_{avg}). We then calculate the difference in complexity between these time spans (δ_C).

We explore whether lexical complexity can explain human uncertainty in annotation by retrieving the human judgments for each pair of sentences for each label for each word (29.000 judgments total), including the label 0 (cannot decide) and compare the complexity difference δ_C between the sentences. We rank the labels by absolute average difference $|\delta_C|$ from largest to smallest, with rank 1 being the label with the highest absolute difference in complexity.

For lexical semantic change detection, we calculate Spearman's rank correlation coefficient between the words' graded lexical change score and δ_C. As baseline, we use a vanilla RoBERTa-base model that was not fine-tuned.

[1]Preliminary experiments have shown a worse performance when using RoBERTa-large and XLM-RoBERTa.

4 Results and Discussion

Table 1 shows the results for the fine-tuned model on the task of lexical complexity prediction. For the limited scope of the study, our model shows acceptable performance. Mean Squared Error (MSE) measures the average deviance from the target, while $R2$ measures the proportion of the variance in the data the model explains. A lower MSE and a higher R^2 are generally better.

	MSE	R^2
Our model (val)	0.0070	0.687
Our model (test)	0.0078	0.524
Best model 2021	0.0061	0.621

Table 1: Results for lexical complexity prediction

Figure 2 shows the clustered column chart for the labels (on the x-axis) and rank counts (on the y-axis) for human uncertainty estimation. The figure clearly shows that the label 0 is ranked first in the majority of cases, indicating that a higher complexity difference coincides with human "cannot decide" judgments. Conversely, label 4 is systematically ranked last, indicating that sentence pairs with low complexity differences are annotated as having the same sense. We can also observe a systematic linear decrease in rank counts for labels 1 down to 3, suggesting that complexity difference inversely correlates with semantic relatedness: the higher the complexity difference, the less probable it is that the word senses in the two sentences are related.

	Spearman's ρ
Baseline	0.077
Our model C^{t1}_{avg}	0.014
Out model C^{t2}_{avg}	-0.089
Our model δ_C	0.444
Best model 2020	0.422
Cassotti et al. 2023	0.757

Table 2: Results for graded lexical semantic change detection

Table 2 shows the results for lexical semantic change detection. As can be gathered from the results, lexical complexity prediction in itself does not correlate with graded semantic change ('Our model' C^{t1}_{avg} and C^{t2}_{avg}), but the difference in lexical complexity ('Our model' δ_C) shows a moderate

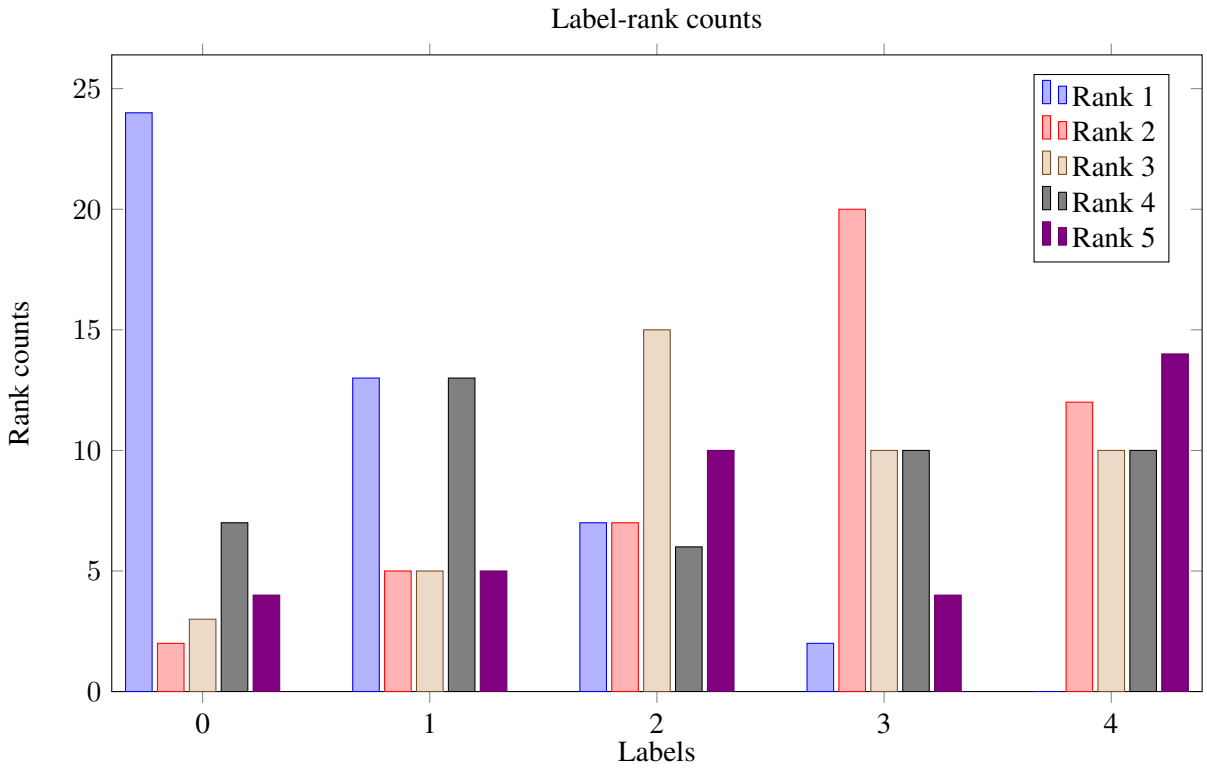

Figure 2: Label-rank counts showing the distribution of rank counts based on complexity difference between the sentence pairs to annotate. We calculate for each label of each word the average complexity difference and rank the labels according to the complexity difference, then aggregate the ranks over all words.

correlation with graded lexical semantic change as calculated by the SemEval 2020 shared task organizers. In fact, this model beats the best result of the shared task, although only by a small margin, and it is quite far behind the current state-of-the-art.

This finding suggests that variations in lexical complexity may be indicative of shifts in word meaning over time.

5 Implications and future work

The findings of this study underscore the interconnected nature of lexical change, highlighting the potential for leveraging lexical complexity prediction in detecting semantic shifts.

Leveraging state-of-the-art machine learning models, such as transformer architectures and contextual embeddings, can enhance the accuracy and scalability of lexical complexity prediction and semantic shift detection.

Our model exhibits surprising performance in graded lexical semantic change detection, outperforming the best result of the shared task by a small margin. While our model's performance would have been competitive, it falls short of the current state-of-the-art models in the field.

In the future, one should extend the analysis to (at least) the other languages covered by the lexical semantic change data (German, Swedish, Latin). However, there are no suitable lexical complexity data sets available for these languages. Hence, it would be necessary to first compile graded lexical resources including different word contexts.

Another promising avenue would be a hybridization of approaches that include lexical complexity prediction as a feature for lexical semantic change detection.

6 Conclusion

In conclusion, our proof-of-concept exploration has shed some light on the interplay between lexical complexity and semantic shift. In this paper, we have shown that pairs of sentences for which the absolute difference in lexical complexity is high tend to be annotated as "cannot decide" by human annotators; this finding suggests that high lexical complexity differences might create ambiguity for human judges, making it difficult for them to confidently discern the exact meaning of a word in the given sentences. We also uncovered a potential inverse correlation between lexical complexity

and semantic relatedness. Finally, we have shown that lexical complexity prediction can be useful for lexical semantic change detection; differences in lexical complexity correlate with graded lexical semantic change to a moderate degree.

Limitations

The presented work focuses on English only due to the availability of resources. It would be beneficial to extend it to other languages. However, results may also be skewed due to the fact that the lexical complexity prediction data set only contains nouns, and the lexical semantic change data set mostly contains nouns. Results might thus not scale to other part-of-speech categories. Further studies with diverse data and evaluation settings are crucial to establish broader validity and generalizability.

Our method shows promising results, but we cannot be sure that it is indeed capturing differences in meaning as expressed through the different word contexts, or whether the model is relying on other (potentially confounding) information.

As a proof of concept study, we only fine-tuned a single model. Future work should explore a wider variety of models. However, fine-tuning models can be costly and may require the use of GPUs. We have only fine-tuned a single (smaller) model, as opposed to a larger or multiple models.

The current work utilizes a relatively limited data set. Therefore results should be interpreted with this limitation in mind.

Acknowledgements

This work has in part been funded by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021).

References

Raksha Agarwal and Niladri Chatterjee. 2021. Gradient Boosted Trees for Identification of Complex Words in Context. In *Proceedings of the First Workshop on Current Trends in Text Simplification*.

David Alfter. 2021. *Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective*. Ph.D. thesis, University of Gothenburg.

David Alfter and Elena Volodina. 2018. Towards Single Word Lexical Complexity Prediction. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 79–88.

Pierpaolo Basile, Annalina Caputo, Tommaso Caselli, Pierluigi Cassotti, and Rossella Varvara. 2020. DIACR-Ita@ EVALITA2020: Overview of the evalita2020 diachronic lexical semantics (diacr-ita) task. *Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. XL-LEXEME: WiC Pretrained Model for Cross-Lingual LEXical sEMantic changE. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585.

Guillaume Chiron, Antoine Doucet, Mickaël Coustaty, and Jean-Philippe Moreux. 2017. ICDAR2017 competition on post-OCR text correction. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1423–1428. IEEE.

Jin-Woo Chung, Hye-Jin Min, Joonyeob Kim, and Jong C Park. 2013. Enhancing readability of web documents by text augmentation for deaf people. In *Proceedings of the 3rd International Conference on Web Intelligence, Mining and Semantics*, pages 1–10.

Scott Crossley, Tom Salsbury, and Danielle McNamara. 2010. The development of polysemy and frequency use in English second language speakers. *Language Learning*, 60(3):573–605.

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. ACM; New York.

Louise Deléger and Pierre Zweigenbaum. 2009. Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora (BUCC)*, pages 2–10.

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

Núria Gala, Thomas François, and Cédrick Fairon. 2013. Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. *E-lexicography in the 21st century: thinking outside the paper., Tallin, Estonia*.

Sian Gooding and Ekaterina Kochmar. 2018. CAMB at CWI Shared Task 2018: Complex Word Identification with Ensemble-Based Voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of

semantic change in the Google Books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK. Association for Computational Linguistics.

Andrey Kutuzov and Lidia Pivovarova. 2021. RuShiftEval: a shared task on semantic shift detection for Russian. *Computational linguistics and intellectual technologies:*.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601, Avignon, France. Association for Computational Linguistics.

Lynn M LoPucki. 2014. System and method for enhancing comprehension and readability of legal text. US Patent 8,794,972.

Tomas Mikolov and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*.

Sara Morsy and George Karypis. 2016. Accounting for language changes over time in document similarity search. *ACM Transactions on Information Systems (TOIS)*, 35(1):1–26.

Alejandro Mosquera. 2021. Alejandro Mosquera at SemEval-2021 Task 1: Exploring Sentence and Word Features for Lexical Complexity Prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 554–559.

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

Jenny A Ortiz-Zambranoa and Arturo Montejo-Ráezb. 2020. Overview of ALexS 2020: First Workshop on Lexical Analysis at SEPLN.

Gustavo Paetzold and Lucia Specia. 2016a. SemEval 2016 Task 11: Complex Word Identification. In *SemEval at NAACL-HLT*, pages 560–569.

Gustavo Paetzold and Lucia Specia. 2016b. Sv000gg at SemEval-2016 task 11: Heavy gauge complex word identification with system voting. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 969–974.

Gustavo Henrique Paetzold. 2021. UTFPR at SemEval-2021 task 1: Complexity prediction by combining BERT vectors and classic features. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 617–622, Online. Association for Computational Linguistics.

Chunguang Pan, Bingyan Song, Shengguang Wang, and Zhipeng Luo. 2021. DeepBlueAI at SemEval-2021 task 1: Lexical complexity prediction with a deep ensemble approach. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 578–584.

Sarah E Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Workshop on Speech and Language Technology in Education*.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111, Athens, Greece. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. Semeval-2020 task 1: Unsupervised lexical semantic change detection. *arXiv preprint arXiv:2007.11464*.

Dominik Schlechtweg, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray. 2021. DWUG: A large Resource of Diachronic Word Usage Graphs in Four Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7079–7091, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthew Shardlow, Richard Evans, Gustavo 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.

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.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2021. Survey of computational approaches to lexical semantic change detection. *Computational approaches to semantic change*, 6(1).

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2004. Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, pages 1385–1392.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jing Wang, Mohit Bansal, Kevin Gimpel, Brian D. Ziebart, and Clement T. Yu. 2015. A Sense-Topic

Model for Word Sense Induction with Unsupervised Data Enrichment. *Transactions of the Association for Computational Linguistics*, 3:59–71.

Tuqa Bani Yaseen, Qusai Ismail, Sarah Al-Omari, Eslam Al-Sobh, and Malak Abdullah. 2021. JUST-BLUE at SemEval-2021 Task 1: Predicting Lexical Complexity using BERT and RoBERTa Pretrained Language Models. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 661–666.

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, United States. Association for Computational Linguistics.

Frank D Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. *LChange 2022*, page 149.