TUDA-CCL at SemEval-2021 Task 1: Using Gradient-boosted Regression Tree Ensembles Trained on a Heterogeneous Feature Set for Predicting Lexical Complexity

Sebastian Gombert and Sabine Bartsch
Corpus and Computational Linguistics, English Philology
Institute of Linguistics and Literary Studies
Technische Universität Darmstadt, Germany
sebastiang@outlook.de, sabine.bartsch@tu-darmstadt.de

Abstract

In this paper, we present our systems submitted to *SemEval-2021 Task 1* on *lexical complexity prediction* (Shardlow et al., 2021a). The aim of this shared task was to create systems able to predict the *lexical complexity* of word tokens and bigram multiword expressions within a given sentence context, a continuous value indicating the difficulty in understanding a respective utterance. Our approach relies on gradient boosted regression tree ensembles fitted using a heterogeneous feature set combining linguistic features, static and contextualized word embeddings, psycholinguistic norm lexica, *WordNet*, word- and character bigram frequencies and inclusion in word lists to create a model able to assign a word or multiword expression a context-dependent complexity score. We can show that especially contextualised string embeddings (Akbik et al., 2018) can help with predicting lexical complexity.

1 Introduction

In this paper, we present our contribution to *SemEval-2021 Shared Task 1* (Shardlow et al., 2021a), a shared task focused on the topic of lexical complexity prediction. The term *lexical complexity prediction* describes the task of assigning a word or multiword expression a continuous or discrete score signifying its likeliness of being understood well within a given context, especially by a non-native speaker. Solving this task could benefit second-language learners and non-native speakers in various ways. One could imagine using such scores to extract vocabulary lists appropriate for a learner level from corpora and literature (Alfter and Volodina, 2018), to judge if a given piece of literature fits a learner's skill or to assist authors of textbooks in finding a level of textual difficulty appropriate for a target audience.

Predicting these scores can be formulated as a regression problem. Our approach to solve this problem relies on *gradient-boosted regression tree ensembles* which we fit on a heterogeneous feature set including different word embedding models, linguistic features, *WordNet* features, psycholinguistic lexica, corpus-based word frequencies and word lists. We assumed that lexical complexity could be correlated with a wide range of features, neural ones as much as distributional or psycholinguistic ones, which is why we chose to use an ensemble-based method in the form of gradient boosting (Mason et al., 1999) for our system as it usually performs best for tasks where such a feature set is needed compared to solely neural models which need dense, homogeneous input data to perform well.

Out of all participants, our systems were ranked *15/54* in the single word- and *19/37* in the multiword category during the official shared task evaluations according to *Pearson's correlation coefficient*. Our key discovery is that while features from nearly all categories provided by us were used by our systems, *contextual string embeddings* (Akbik et al., 2018) were the by far most important category of features to determine lexical complexity for both systems. The code and our full results can be found at https://github.com/SGombert/tudacclsemeval.

2 Background

2.1 Task Setup

For the shared task, *CompLex* corpus (Shardlow et al., 2020, 2021b) was used as data set. This English corpus consists of sentences extracted from the *World English Bible* of the multilingual corpus consisting of bible translations published by Christodoulopoulos and Steedman (2015), the English version of *Europarl* (Koehn, 2005), a corpus containing various texts concerned with European policy, and *CRAFT* (Bada et al., 2012), a corpus

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 130–137
Bangkok, Thailand (online), August 5–6, 2021.

consisting of biomedical articles.

CompLex is divided into two sub-corpora, one dealing with the complexity of single words and the other one with the complexity of bigram multiword expressions. Accordingly, the shared task was divided into two sub-tasks, one dedicated to each sub-corpus. Within both *CompLex* sub-corpora, the sentences are organised into quadruples consisting of a given sentence, a reference to its original corpus, a selected word, respectively a multiword expression from this sentence, and a continuous complexity score denoting the difficulty of this selected word or bigram which is to be predicted by systems submitted to the shared task. For the task, both subcorpora were partitioned into training, test and trial sets.

The scores given for simple words, respectively multiword expressions, were derived from letting annotators subjectively judge the difficulty of understanding words respectively word bigrams on a Likert scale ranging from *1* to *5* with *1* indicating a very simple and *5* a very complex word. The assigned scores were then projected onto values between *0* and *1* and averaged between all annotators to calculate the final scores.

2.2 Related Work

The first approaches to the systematic prediction of lexical complexity were made during *SemEval-2016 Task 11* (Paetzold and Specia, 2016). Here, the problem of determining the complexity of a word was formulated as a classification task designed to determine whether a word could be considered as being complex or not. The data set used for this task was created by presenting 20 non-native speakers with sentences and letting them judge whether the words contained within these sentences were rated as complex or not. From these judgements, two different data sets were derived. In the first one, a word was considered complex if at least one of the annotators had judged it as such, and in the second one, each word was given 20 different labels, one per annotator. The most important findings for this shared task were that ensemble methods performed best in predicting lexical complexity with word frequency being the best indicator.

In 2018, a second shared task was conducted on the same topic as described in Yimam et al. (2018). This shared task focused on predicting lexical complexity for English, German, Spanish and a multilingual data set with a French test set. The data for this was acquired by presenting annotators on *Amazon Mechanical Turk* with paragraphs of text and letting them mark words which according to their perception could hinder the same paragraph from being understood by a less proficient reader. The findings of this shared task confirmed the finding of the previous one that using ensemble methods yield best results for complex word identification with a system submitted by Gooding and Kochmar (2018) relying on decision tree ensembles.

3 System Overview

Our systems rely on *gradient-boosted regression tree ensembles* (Mason et al., 1999) for predicting lexical complexity scores. We trained one model to predict single word lexical complexity scores and another one to predict bigram multiword expression complexity scores. Our models are based on the implementation of gradient boosting provided by *CatBoost*[1] (Dorogush et al., 2018; Prokhorenkova et al., 2018). We set the growing policy to *loss-guide*, the *L2 leaf regularisation* to *15*, the *learning rate* to *0.01*, *tree depth* to *6* and the *maximum number of leaves* to 15. Additionally, we set the *number of maximum iterations* to *5000* and then used the trial set to perform *early stopping* during training in order to determine the exact number of required iterations.

The motivation behind using this algorithm was its general ability to perform well on heterogeneous and sparse feature sets which allowed us to mix regular linguistic features, *WordNet* features, word embeddings, psycho-linguistic norm lexica, corpus-based word frequencies and selected word lists as all of these were features we assumed to possibly correlate with lexical complexity. Moreover, the reportings of Paetzold and Specia (2016) and Yimam et al. (2018) that ensemble-based learners perform best for *complex word identification* contributed to this decision, as well. While the problem presented in their paper is formulated as a binary classification task using different data sets, we wanted to test if their findings would still translate to a regression task on *CompLex*.

3.1 Feature Engineering

The following paragraphs describe the features we used to create the feature vectors used to represent words. In case of our system dealing with bigram

[1] https://catboost.ai/

multiword expressions, we calculated such a vector for each of both words and then concatenated them to acquire the final input vectors. Thus, the exact number of input features was *7424* for our system dealing with single words and *14848* for our system dealing with multiword expressions.

Syntactic features: This category of features includes *XPOS*-, *UPOS*-, *dependency*- and *named entity tags* as well as *universal features*[2] inferred using the English *Stanza*[3] (Qi et al., 2020) model fit to the version of the *English Web Treebank* following the *Universal Dependencies* formalism (Silveira et al., 2014). In addition to the tags assigned to the word(s) whose score was to be predicted, we included the *XPOS*- and *UPOS* tags of the two neighbouring words to the left and to the right as well as the dependency tags of the siblings, direct children and the parent of the word(s) within the dependency structure of a given sentence. All of these features are encoded as *one*-, respectively *n-hot vectors* using the *LabelBinarizer* and *MultiLabelBinarizer* classes provided by *Scikit-learn* (Pedregosa et al., 2011).

WordNet features: Here, we included the numbers of *hypernyms*, *root hypernyms*, *hyponyms*, *member holonyms*, *part meronyms* and *member meronyms* of the respective word(s) as well as the number of given examples and the *length of the shortest hypernym path* from *WordNet* (Miller, 1995). In cases where multiple *synsets* were given for a word, we calculated the respective means and in cases where a given word was not included in the resource, we set all respective feature values to *0*. We accessed *WordNet* using *NLTK* (Bird et al., 2009). The main intuition behind using this resource was that the *length of the shortest hypernym path* and the count for the different lexico-semantic relations could be a good indicator for lexical complexity.

Word embeddings: We used multiple static and contextual word embedding models for our feature set. This includes the *transformer-based* (Devlin et al., 2019) *BiomedNLP-PubMedBERT-base-uncased-abstract* (Gu et al., 2020), *distilgpt2*[4] (Radford et al., 2018) and *distilbert-base-uncased* (Sanh et al., 2019), the *contextual string embedding* models *mix-forward* and *mix-backward*[5] (Akbik et al., 2018), and the static *GloVe*[6] (Pennington et al., 2014) and English *fastText*[7] (Bojanowski et al., 2017) embeddings.

This collection of embeddings was derived from previous experiments on the *CompLex* corpus where we tried to fine-tune a purely neural model using the approach of stacking different embedding models in combination with an attached prediction head central to *flairNLP*[8] (Akbik et al., 2019). More precisely, in the setup we chose, the outputs of all language models were fed to a feed-forward layer responsible for calculating the final complexity scores. This network was then trained for *5* epochs with a *learning rate of 0.000001*, *mean squared error* as loss function and *Adam* (Kingma and Ba, 2015) as optimizer on the training set part of *CompLex*. During this training, fine-tuning was active for all transformer-based language models so that their weights were adjusted during the process and *scalar mixing* (Liu et al., 2019) was used for the transformer-based language models as it was not foreseeable which layers of the transformer models would influence results the most.

This model achieved a *Pearson's correlation coefficient* score of 0.7103 when evaluated on the trial set. While we deemed this an okay result, we decided to stick with gradient boosting for our final systems as early experiments with this algorithm yielded results superior to the purely neural approach when evaluated on the same set. As we switched to using gradient boosting for our final systems, we decided to use the fine-tuned variants of the transformer embedding models as using them led to small improvements when testing our models on the shared task trial sets compared to using the non-fine-tuned variants.

Psycholinguistic norm lexica: Our feature set includes two psycholinguistic norm lexica. The first one is described in Warriner et al. (2013) and scores words with empirical ratings for *pleasantness*, *arousal* and *dominance* using the *SAM score* (Bradley and Lang, 1994). These ratings were acquired from annotators on the *Amazon Mechanical Turk* platform. The second lexicon is described in

[2] https://universaldependencies.org/u/feat/all.html
[3] https://stanfordnlp.github.io/stanza/
[4] https://huggingface.co/distilgpt2
[5] https://github.com/flairNLP/flair/blob/master/resources/docs/embeddings/FLAIR_EMBEDDINGS.md
[6] https://nlp.stanford.edu/projects/glove/
[7] https://fasttext.cc/
[8] https://github.com/flairNLP/flair

Malandrakis and Narayanan (2015) and includes ratings for *arousal*, *dominance*, *valence*, *pleasantness*, *concreteness*, *imagability*, *age of acquisition*, *familarity*, *pronouncability*, *context availability* and *gender ladenness*. The ratings within this lexicon were derived algorithmically from smaller lexicons using linear combinations and semantic similarity scores to approximate the ratings for words not included in the source lexica. In both cases, the inclusion of these features was mainly motivated by our general intuition that the perceived complexity of words could be linked to different psycholinguistic variables.

Word frequencies: We utilised three resources containing corpus-based word respectively character bigram frequencies. The first of these data sets was the frequency list extracted from the *SUBTLEXus* corpus (Brysbaert and New, 2009) consisting of various movie subtitles from which we used the *log-normalised* term frequency and the *log-normalised* document frequency as features. Besides *SUBTLEXus*, we utilised the character bigram frequencies from Norvig (2013) which were extracted from the *Google Books Corpus*. Here, to represent a word, we calculated the mean of all frequencies of the bigrams consituting the same and used this as feature. In the case of both sets, our intuition was that lower frequency would likely function as a proxy for complexity. The third set we used was *EFLLex* (Dürlich and François, 2018) which lists the frequencies of words within several pieces of English literature appropriate for different *CEFR*[9] levels. We included this set as we deemed that *CEFR* as a framework for rating language competence could also function as an according proxy.

Word Lists: We used two different word lists as features. The first one is *Ogden's Basic English Vocabulary*[10], a list of simple words used for writing *simple English* as described in Ogden (1932). Here, our idea was that this could help to identify simple words within *CompLex*. The other one was the *Academic Word List* as described in Coxhead (2011), a structured lexicon of terms used primarily in academic discourse which we believed to contain more complex words. In both cases, we encoded the inclusion of a word within a respective word list binarily.

[9] https://tracktest.eu/english-levels-cefr/

[10] http://ogden.basic-english.org/

Metric	System	Rank	Best Res.
Pearson	0.7618	15/54	*0.7886*
Spearman	0.7164	26/54	*0.7425*
MAE	0.0643	20/54	*0.0609*
MSE	0.0067	9/54	*0.0061*
R2	0.5846	10/54	*0.6210*

Table 1: Results achieved by our system dealing with single word complexity. **Best Res**ults refer to the best score achieved within each category by a competing system.

Metric	System	Rank	Best Res.
Pearson	0.8190	19/37	*0.8612*
Spearman	0.8091	19/37	*0.8548*
MAE	0.0711	14/37	*0.0616*
MSE	0.0080	12/37	*0.0063*
R2	0.6677	13/37	*0.7389*

Table 2: Results achieved by our system dealing with multiword expression complexity. **Best Res**ults refer to the best score achieved within each category by a competing system.

4 Results

Throughout the shared task, the systems were evaluated with regard to *Pearson's correlation coefficient*, *Spearman's rank correlation coefficient*, *mean average error*, *mean squared error* and *R2* with *Pearson's correlation coefficient* determining the main ranking. According to this, our systems achieved the 15th and 19th rank respectively. Table 1 shows the results achieved by our system dealing with single words and Table 2 the results achieved by our system dealing with multiword expressions. The results show that our systems, while only achieving upper mid-table results on average, come close to the best systems performance-wise which speaks for our approach. Further hyperparameter tuning and the addition of more features could likely close this gap. The full results for all submitted systems are presented in Shardlow et al. (2021a).

4.1 Most Important Features

To determine which features were used by our models to predict lexical complexity, we rely on the functionality provided by *CatBoost* which scores each feature for its influence on a given final prediction. This is achieved by changing a respective feature values and observing the resulting change

Rank	Feature	Importance
1	flair-mix-b.	25.10
2	flair-mix-b.	11.79
3	flair-mix-b.	7.03
4	flair-mix-f.	4.09
5	flair-mix-f.	2.98
6	flair-mix-b.	1.33
7	flair-mix-f.	1.20
8	distilbert-b.-u.	1.19
9	BiomedNLP	1.12
10	GloVe	1.03

Table 3: The 10 most important features observed for our system dealing with single word complexity and their categories. Each entry refers to a single dimension of the feature vector.

Rank	Feature	Importance
1	flair-mix-b. (2nd w.)	9.28
2	flair-mix-b. (2nd w.)	7.24
3	flair-mix-b. (2nd w.)	6.09
4	flair-mix-b. (2nd w.)	3.80
5	flair-mix-b. (2nd w.)	3.60
6	flair-mix-b. (2nd w.)	3.17
7	flair-mix-b. (2nd w.)	2.44
8	flair-mix-b. (2nd w.)	1.88
9	flair-mix-b. (2nd w.)	1.34
10	flair-mix-b. (2nd w.)	1.08

Table 4: The 10 most important features observed for our system dealing with multiword expression complexity and their categories. Each entry refers to a single dimension of the feature vector.

in the model prediction (see [11] for further information on the exact method). The outputs of this method are normalised so that the sum of the importance values of all features equals *100*. *Feature importance* was calculated using the evaluation set of *CompLex*.

Inspecting the results of these calculations, we noticed that our systems did not use the character bigram frequencies derived from the *Google Books Corpus*, nor the frequencies from *EFLLex* or the word list inclusion features. While features from all other categories were utilised, the most dominant features by far are contained in the word embedding category. Within this category, the most dominant features for both models came from the *flair-mix-backward* and *flair-mix-forward* models (see Tables 3 and 4). A few single dimension from the embeddings provided by *flair-mix-backward* seem to play the major role here.

In the case of our model dealing with multiword expressions, the ten most important features all stem from the *flair-mix-backward* embedding of the second word. This could be explained by the fact that most multiword expressions within the *CompLex* corpus follow the structure of a semantic head in combination with a modifier as most of them are either multi token compounds or single token nouns modified by adjectives. It is intuitive from a linguistic point of view that in such cases, the semantic head, which comes as second element, should play the dominant semantic role resulting in it being more influential in the overall results.

[11] https://catboost.ai/docs/concepts/fstr.html

While the exact reason for the strong influence of the *contextualised string embeddings* is hard to determine due to the fact that embeddings lack the property of being easily interpretable, we assume that the dominant role they play for the results could be determined by them being calculated on the character level (Akbik et al., 2018) instead of the level of fixed words or subword units such as morphemes. As a consequence, such models use fewer input dimensions and each of the dimensions present is in turn involved in the encoding of more different words. This links each input dimension also to a larger variety of latently encoded distributional knowledge which could then contain certain regularities strongly correlated with lexical complexity. However, without further research, this currently remains pure speculation.

4.2 Predictions vs. Ground Truth

In order to compare the predicted values of our models to the ground truth data, we scatterplotted the relationship between ground truth labels and the scores predicted by our systems (see Figures 1 and 2) using the *CompLex* evaluation set. It can be observed that both systems, especially the one dealing with single word complexity, show the tendency to assign slightly higher scores than given in the ground truth for simple words and slightly lower scores for complex words. The system dealing with multiword expressions does not assign any value below 0.2 at all and the one dealing with single word complexity rarely does so. This indicates that our feature set does not contain features which could help our models to identify very simple words.

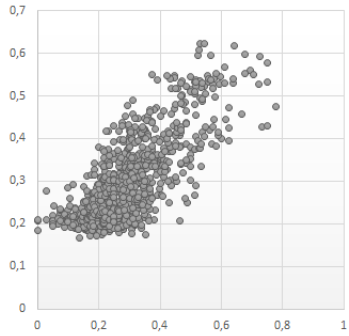

Figure 1: Scatterplot visualising the relationship between the ground truth and the predictions of our model for single word complexity. **X:** ground truth **Y:** prediction

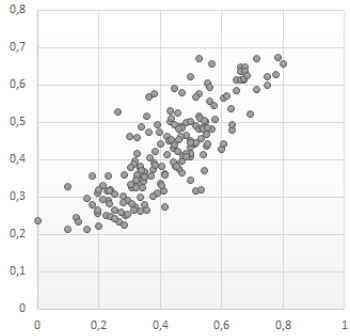

Figure 2: Scatterplot visualising the relationship between the ground truth and the predictions of our model for multiword expression complexity. **X:** ground truth **Y:** prediction

5 Conclusion

We presented both our systems submitted to *SemEval-2021 Task 1* combining a heterogeneous feature set with gradient boosting as regression algorithm. Our systems ware ranked *15/54* and *19/37* during shared task evaluations according to *Pearson's correlation coefficient*. However, the results achieved by our systems were still close to the best results, especially in the case of the system dealing with single word complexity. The type of feature playing the most important role for our models are *contextual string embeddings* as they influenced the outcome the most. We attribute this to a relationship between lexical complexity and the distribution of characters throughout words and sentences, but this needs further clarification which could be the objective of future work. Moreover, our systems rarely assign scores below 0.2. It must be explored further if there are features which could improve our systems in this respect. In summary, we can report that ensemble methods turned out to be fruitful when applied to *CompLex*.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59, Minneapolis, Minnesota. Association for Computational Linguistics.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

David Alfter and Elena Volodina. 2018. Towards single word lexical complexity prediction. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 79–88, New Orleans, Louisiana. Association for Computational Linguistics.

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William Baumgartner Jr, Kevin Cohen, Karin Verspoor, Judith Blake, and Lawrence Hunter. 2012. Concept annotation in the CRAFT corpus. *BMC bioinformatics*, 13:161.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59.

Marc Brysbaert and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4):977–990.

Christos Christodoulopoulos and Mark Steedman. 2015. A massively parallel corpus: the Bible in 100 languages. *Lang. Resour. Evaluation*, 49(2):375–395.

Averil Coxhead. 2011. The academic word list 10 years on: Research and teaching implications. *TESOL Quarterly*, 45(2):355–362.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. CatBoost: gradient boosting with categorical features support. *CoRR*, abs/1810.11363.

Luise Dürlich and Thomas François. 2018. EFLLex: A graded lexical resource for learners of English as a foreign language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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, New Orleans, Louisiana. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Nikolaos Malandrakis and Shrikanth S. Narayanan. 2015. Therapy language analysis using automatically generated psycholinguistic norms. In *Proceedings of Interspeech*.

Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. 1999. Boosting algorithms as gradient descent. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, page 512–518, Cambridge, MA, USA. MIT Press.

George A. Miller. 1995. WordNet: A lexical database for English. *Commun. ACM*, 38(11):39–41.

Peter Norvig. 2013. English letter frequency counts: Mayzner revisited or ETAOIN SRHLDCU.

C.K. Ogden. 1932. *Basic English: A General Introduction with Rules and Grammar*. Psyche miniatures. K. Paul, Trench, Trubner & Company, Limited.

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.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 6639–6649, Red Hook, NY, USA. Curran Associates Inc.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language models are unsupervised multitask learners.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

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

Matthew Shardlow, Richard Evans, Gustavo Paetzold, and Marcos Zampieri. 2021a. Semeval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2021)*.

Matthew Shardlow, Richard Evans, and Marcos Zampieri. 2021b. Predicting lexical complexity in English texts. *arXiv preprint arXiv:2102.08773*.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

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.