To What Extent Can English-as-a-Second Language Learners Read Economic News Texts?

Yo Ehara
Tokyo Gakugei University
4-1-1 Nukuikita-machi, Koganei-shi, Tokyo 184-8501 Japan
ehara@u-gakugei.ac.jp

Abstract

In decision making in the economic field, an especially important requirement is to rapidly understand news to absorb ever-changing economic situations. Given that most economic news is written in English, the ability to read such information without waiting for a translation is particularly valuable in economics in contrast to other fields. In consideration of this issue, this research investigated the extent to which non-native English speakers are able to read economic news to make decisions accordingly – an issue that has been rarely addressed in previous studies. Using an existing standard dataset as training data, we created a classifier that automatically evaluates the readability of text with high accuracy for English learners. Our assessment of the readability of an economic news corpus revealed that most news texts can be read by intermediate English learners. We also found that in some cases, readability varies considerably depending on the knowledge of certain words specific to the economic field.

1 Introduction

In the economic field, it is important to read news as primary information to make decisions quickly. While the majority of economic news is written in English, the economic activities of non-native English speakers, or English-as-a-Second-Language (ESL) learners also have a strong influence. To what extent can ESL learners read economic news? If, for example, certain economic news is difficult for ESL learners to read, then the language gap may affect their economic activities. Although this question is important, it was rarely addressed in previous studies.

The importance of the aforementioned question is also related to the *efficient market hypothesis*. The hypothesis is one of the major propositions in financial economics that forms the basis of the Black–Scholes equation, which is a fundamental formula used for market modeling. It assumes that "asset prices reflect all available information" [1]. Intuitively, the hypothesis maintains that no information gap, including the gaps caused by language deficiencies, exists in the market. However, this proposition does not hold in cases wherein most second language learners cannot read economic news to make economic decisions in the market. We tackled the challenge of illuminating this issue.

To this end, we first sought to understand the difficulty that economic news poses for non-native speakers of English. We approached this problem in two ways. The first was based on the field of educational NLP (Vajjala and Lučić, 2018). Using a corpus that is standard in this field, we constructed a machine learning classifier that can determine the difficulty of a text with high accuracy using deep learning methods, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019).

The second approach was to conduct readability assessments on the basis of information about the vocabulary of English learners. These methods have been well studied in the field of applied linguistics, where considerable research has shown that English learners need to know more than 95% of words in a text to read and understand them (Nation, 2006; Laufer and Ravenhorst-Kalovski, 2010). The idea of assessing text readability via each learner's vocabulary knowledge is beneficial for interpreting readability assessment results. Therefore, we also constructed a classifier that ascertains how many words in a text an English learner knows using a data set of English learners' vocabulary tests (Ehara, 2018).

In experiments carried out on a standard data set for evaluating readability (Vajjala and Lučić, 2018) in educational NLP, the two approaches to assessing readability derived results that were in close

[1] https://en.wikipedia.org/wiki/Efficient-market_hypothesis

agreement. Next, the experiments involving a real economic news corpus uncovered that most of the texts in the economic news corpus were readable by intermediate English learners. The analysis for which the second approach was used indicated that readability can be substantially improved for English learners through knowledge of a few words that are frequently used in economics.

The contributions of this study are as follows:

1. We constructed a high-performance readability evaluator and examined the readability of economic news for English learners.

2. We showed that intermediate English learners can read economic news texts.

3. We demonstrated that the knowledge of a few economic domain-specific words may considerably improve the readability of some economic news for second language learners.

2 Automatic Readability Assessment

This section formalizes the problem of automatic readability assessment. Let us suppose that we have N texts to assess: we write the set of texts as $\{\mathcal{T}_i | i \in \{1, \ldots, N\}\}$. Let $\mathcal{Y}$ be the set of readability labels. Labels are typically ordered in the order of difficulty. For example, in the *OneStopEnglish* dataset (Vajjala and Lučić, 2018), we can set $\mathcal{Y} = \{0, 1, 2\}$, where 0 is elementary, 1 is intermediate, and 2 is advanced. The number of levels depends on the evaluation corpus. Using $\mathcal{Y}$, we write the label for $\mathcal{T}_i$ as $y_i \in \mathcal{Y}$.

Given each text $\mathcal{T}_i$, an *assessor* outputs its readability score s_i. In a supervised setting, the *assessor* knows the number of levels in the evaluation corpus from training examples. Hence, s_i ranges within $\mathcal{Y}$: $s_i \in \mathcal{Y}$. However, in an unsupervised setting, it is noteworthy that the assessor does not know $\mathcal{Y}$, or how many levels the evaluation corpus has, because no label is given. Hence, even if only integers are allowed for y_i, s_i can be a real value.

Throughout this paper, we write arrays using [and]. Given N texts $[\mathcal{T}_i | i \in \{1, \ldots, N\}]$, our goal is to make an assessor output arrays of readability scores $[s_i | i \in \{1, \ldots, N\}]$ that *correlate well* with the array of labels $[y_i | i \in \{1, \ldots, N\}]$. Here, there are multiple types of correlation coefficients between the array of scores and the array of labels, which we explain in the later sections. Typically, we should use *rank coefficients* such as Spearman's ρ, defined as the Pearson's ρ between rankings, when s_i is real-valued.

```
15. deficit:
The company <had a large deficit>.
a: spent a lot more money than it earned
b: went down a lot in value
c: had a plan for its spending
                that used a lot of money
d: had a lot of money stored in the bank
```

Figure 1: Examples of the Vocabulary Size Test, one of the most widely accepted vocabulary tests to quickly assess language learners. They are asked to choose the option that paraphrases the part between "<" and ">" from a, b, c, and d.

3 Vocabulary Testing-based Readability

Fig. 1 shows example questions from the vocabulary size test, a widely used vocabulary test in applied linguistics (Beglar and Nation, 2007). Each question asks about a word in a multiple-choice question format. The test consists of 100 questions like those shown in Fig. 1. Ehara (2018) used this test to have 100 second-language learners take the test and to collect their responses. Their data were published and made publicly available. We used their dataset to train our classifiers.

We want to analyze vocabulary test results to obtain word difficulty values encoding learners' language knowledge. To this end, we employed the idea of *item response theory* (Baker, 2004), a statistical model that can estimate learners' abilities and test questions' difficulties from the learners' responses to the questions.

Let $\mathcal{V}$ be the set of vocabulary, and let $\mathcal{L}$ be the set of learners. Let $z_{v,l} \in \{0, 1\}$ be the result of whether learner $l \in \mathcal{L}$ correctly answered the question for word $v \in \mathcal{V}$: $z_{l,v} = 1$ if l answered correctly for word v; otherwise, $z_{l,v} = 0$. Correct answers usually imply that l knows word v.

Then, by using $\{z_{v,l}\}$ as the training data, we train the following model:

$$p(z = 1 | v, l) = \text{sigmoid}(a_l - d_v) \qquad (1)$$

In Eq. 1, a_l is the ability parameter of learner l, d_v is the difficulty of word w, and sigmoid denotes the logistic sigmoid function, i.e., $\text{sigmoid}(x) = \frac{1}{1+\exp(-x)}$.

The logistic sigmoid function is the binary version of the softmax function, which is frequently used in neural classifiers. It is a monotonously

increasing function ranging within $(0, 1)$. As $\text{sigmoid}(0) = \frac{1}{1+1} = \frac{1}{2}$, when a learner's ability a_l is larger than the word difficulty d_v, the probability that learner l knows word v can be written as follows: $p(z = 1|v, l) > \frac{1}{2}$ in Eq. 1. Likewise, by using Eq. 1, we can compare a learner's ability and word difficulty in the same dimension.

To estimate learner ability and word difficulty, $z_{v,l}$ is given as z in Eq. 1 in the training phase. In this way, in *item response theory*, learner ability and word difficulty are comparable, and these parameters are estimated from the test result data.

In Eq. 1, d_v denotes the word difficulty estimated from the vocabulary tests. Here, in addition to the word difficulty for the words within the vocabulary test, we also want to obtain word difficulty values for all words that may appear in the target language. To this end, we calculate d_v by using the word frequency in large balanced corpora as features as follows:

$$d_v = -\sum_{k=1}^{K} w_k \log(\text{freq}_k(v) + 1) \qquad (2)$$

Eq. 2 assumes that we have K corpora to use as features to calculate word difficulty d_v. In Eq. 2, K is the number of corpora to use, $\text{freq}_k(v)$ denotes the frequency of word v in the k-th corpus, and w_k is the weight parameter of the k-th corpus. In summary, given the vocabulary test results $\{z_{v,l}\}$ and corpus frequency features $\text{freq}_k(v)$, we can estimate the parameters: namely, the weight of the k-th corpus w_k and learner l's ability a_l. To implement the model, we used logistic regression, by following (Ehara, 2018). Note that this model does not use the valuable readability label $\{y_i\}$ in the training phase, so is unsupervised.

After estimating the parameters using the above-mentioned procedure, we use the following formula to obtain the readability of given $\mathcal{T}_i$. Here, l_{avg} denotes the test-taker whose estimated ability parameter is closest to the average of the estimated ability parameter values $\{a_l\}$. Intuitively, the following equation calculates the probability that the average learner knows all the words that appear in $\mathcal{T}_i$ and uses it as the readability score. The use of the $-\log$ here has two reasons. The reason of using $\log$ is to prevent problems in numerical calculations, since the probability values can be very close to 0. While we want to design d_v so that the more difficult the word v, the larger the value, whereas the higher the probability, the easier the word. Hence, we use negative of $\log$ to make the two scales meet.

$$s_i = score(\mathcal{T}_i) = -\log\left(\prod_{v \in \mathcal{T}_i} p(z = 1|v, l_{\text{avg}})\right) \qquad (3)$$

4 Experimental Settings

4.1 Readability Dataset

We used the OneStopEnglish dataset (Vajjala and Lučić, 2018) for the source of readability for second language learners because it is one of the newest, publicly available, and reliable in the sense that no known trivial features are effective for predicting its labels such as average sentence length.

The dataset has three levels: elementary, intermediate, and advanced. The original articles were taken from the Guardian newspaper. The OneStopEnglish dataset is a parallel corpus, i.e, language teachers manually rewrote the original articles into the three aforementioned readability levels. Hence, the corpus is designed so that its readability labels are not easily estimated from the topic of texts. While a text is usually a single newspaper article in the OneStopEnglish dataset. However, during the building of the dataset, it can be edited to be a unit shorter than an article. Therefore, henceforth, we simply call it a text.

All three levels have 189 texts each, 567 texts in total. We split these texts into a *training set* consisting of 339 texts, a *validation* set consisting of 114 texts, and a *test* set consisting of 114 texts. The *training set* and *validation* sets were used to train solely supervised methods for comparison. Unsupervised methods did not use the training and validation sets; they used only the test set.

4.2 Compared Methods

As the BERT-based sequence classification has been reported to achieve excellent results (Devlin et al., 2019), we applied the standard BERT-based sequence classification approach involving pretraining and fine-tuning. For the pretrained model, we used **bert-large-cased-whole-word-masking** in the Huggingface models [2].

Then, we fine-tuned the model using the 339 training texts. We named this fine-tuned model **spvBERT**, in which "spv" denotes being supervised. For fine-tuning, we used the Adam optimizer

[2] https://huggingface.co/models

Method	Spearman's ρ	Pearson's ρ
Flesch-Kincaid	0.324	0.359
ARI	0.317	0.351
Coleman-Liau	0.373	0.372
FleschReadingEase	-0.387	-0.426
GunningFogIndex	0.331	0.362
LIX	0.348	0.383
SMOGIndex	0.456	0.479
RIX	0.437	0.462
DaleChallIndex	0.495	0.506
TCN RSRS-simple	-	0.615(*)
Vocabulary-based	**0.730**	**0.715**
spvBERT	0.866	0.864

Table 1: Predictive Performance of Readability. Only **spvBERT** is supervised: the others are unsupervised.

Year/Month	Elem.	Int.	Adv.
Jan.2007	0.001	0.970	0.029
Feb.2007	0.003	0.975	0.022
Mar.2007	0.006	0.965	0.029
Apr.2007	0.002	0.976	0.022
May 2007	0.001	0.978	0.021
Jun.2007	0.009	0.964	0.027
Jul.2007	0.002	0.964	0.034
Aug.2007	0.002	0.951	0.047
Sep.2007	0.001	0.953	0.046
Oct.2007	0.005	0.949	0.047
Nov.2007	0.001	0.955	0.044
Dec.2007	0.012	0.944	0.044

Table 2: Readability Assessment Results of Economic News Texts in 2007. "Elem." denotes "elementary", "Int." denotes "intermediate", and "Adv." denotes "advanced".

(Kingma and Ba, 2015) with a setting of 10 epochs and a 10^{-5} training rate.

For the implementation of conventional readability formulae, we used the **readability** PyPI package [3]. We used almost all readability formulae implemented in this package for our experiments: namely, **Flesch-Kincaid** (Flesch-Kincaid Grade Level, FKGL) (Kincaid et al., 1975), **ARI** (Automated Readability Index) (Senter and Smith, 1967), the **Coleman-Liau** Index (Coleman and Liau, 1975), **Flesch Reading Ease** (Flesch, 1948), the **Gunning Fog Index** (Gunning, 1952), **LIX** (Björnsson, 1968), the **SMOG Index** (Mc Laughlin, 1969), the RIX index (Anderson, 1983), and the **Dale-Chall Index** (Dale and Chall, 1948). More details of these formulae and their implementation are described on the project page. All of these readability formulae are *unsupervised* in the sense that they do not require any training data.

The **Vocabulary-based** model was trained on a publicly available vocabulary dataset (Ehara, 2018). For the corpus word frequency, we used the frequencies taken from the British National Corpus (BNC Consortium, 2007) and the Corpus of Contemporary American English (COCA) (Davies, 2008). Both corpora are balanced general corpora used extensively in English education (Nation, 2006). Especially, the word frequencies of these corpora are important resources for determining word difficulty in English education.

[3] https://pypi.org/project/readability/

4.3 Experimental Results

Tab. 1 shows the experimental results. First, in all unsupervised methods, **Vocabulary-based** achieved the best results in all rank correlation coefficients. **TCN RSRS-simple** is the best model on the OneStopEnglish dataset in Martinc et al. (2021). As they show only the performance measured by the Pearson correlation, we filled "-" for Spearman's ρ in the table. While a direct comparison is not possible as denoted by $(*)$, **Vocabulary-based** outperforms it.

Importantly, we can observe that both **Vocabulary-based** and **spvBERT** achieve high predictive performance. This result indicates that the two approaches to assessing readability derived results that were in close agreement.

5 Experiments with Economic News

We used an economic news article dataset (Ding et al., 2014, 2015) for our analysis because it is publicly available and easy to replicate. The dataset consisted of 109,110 Reuters news articles published in 2006–2007. Tab. 2 shows the readability assessment results for each month of 2007. Each month included approximately 1,000 articles. Few of the articles were elementary, whereas most were intermediate. Approximately 2%–4% of the texts were advanced. The results show that intermediate English learners could read most but not 2%–4% of economic news.

We also conducted a vocabulary-based analysis. Using (Ehara, 2018), we chose a learner who

could successfully answer 75 vocabulary questions among the 100 questions in the Vocabulary Size Test (Beglar and Nation, 2007). This learner was estimated to know 15,000 words. Knowledge of this number of words should be sufficient for reading typical newspaper texts (Nation, 2006). However, in the vocabulary-based readability assessor, some words, such as "annuity" and "veritable", were predicted to be unfamiliar to this learner. In one article, these words greatly reduced the probability that this learner could read the text from 0.62 to 0.38: these probability values represented the probability that this learner knew at least 95% of the words in this text. The probability of knowing each word was represented by Eq. 1. Ehara (2019) proposed an algorithm to calculate the probability of knowing 95% or more of the words in a text from the probability of knowing each word. We simply used their algorithm to calculate these probabilities.

6 Discussion

In Eq. 2, word frequencies in the corpora are used as features. In this study, we used multiple balanced corpora. Eq. 2 is applicable when preparing K corpora in advance. k denotes the index of the prepared corpora.

In NLP, the **Proposed** method is closely related to complex word identification (CWI) tasks (Yimam et al., 2018; Paetzold and Specia, 2016). CWI is a task that aims to discover difficult words in a text. The relationship between CWI and personalized text readability was previously studied in (Ehara, 2019). The task of obtaining the difficulty of an English word for each individual ESL learner, as we performed in this study, can be regarded as personalized CWI (Ehara et al., 2012, 2014) [4]. Personalized CWI has many downstream applications in NLP, such as lexical simplification (Lee and Yeung, 2018, 2019), text recommendation for language learners (Ehara et al., 2013; Yeung and Lee, 2018; Lee, 2021), and translator selection in crowdsourcing (Ehara et al., 2016). Some studies have focused on the relationship between word semantics and word difficulty (Ehara et al., 2014; Beinborn et al., 2016; Ehara, 2020b). Regarding the interpretability of CWI classifiers, Ehara (2020a) studied the relationship between CWI classifiers' weights and vocabulary sizes.

[4]The journal version of (Ehara et al., 2012) is (Ehara et al., 2018).

7 Conclusions

In this paper, we focused on the readability of economic news texts for ESL learners. We conducted two approaches for measuring readability: the BERT-based approach and the vocabulary-based approach. We found that although most texts were readable to intermediate learners, 2.4% of articles were not readable to them. Furthermore, some economic words greatly reduced the readability for ESL learners. Future work will include investigating for differences in the readability of different types of economic news for ESL learners.

Acknowledgment

This study was supported by JST ACT-X with Grant Number JPMJAX2006 and JSPS KAKENHI with Grant Number 18K18118. We used the ABCI infrastructure from AIST for the computational resources. We appreciate anonymous reviewers for their valuable comments.

References

Jonathan Anderson. 1983. Lix and rix: Variations on a little-known readability index. *Journal of Reading*, 26(6):490–496.

Frank B. Baker. 2004. *Item Response Theory : Parameter Estimation Techniques, Second Edition*. CRC Press.

David Beglar and Paul Nation. 2007. A vocabulary size test. *The Language Teacher*, 31(7):9–13.

Lisa Beinborn, Torsten Zesch, and Iryna Gurevych. 2016. Predicting the Spelling Difficulty of Words for Language Learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 73–83, San Diego, CA. Association for Computational Linguistics.

C. H. Björnsson. 1968. Läsbarhet, Stockholm.

BNC Consortium. 2007. The british national corpus, version 3 (bnc xml edition). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium http://www.natcorp.ox.ac.uk/.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Mark Davies. 2008. The corpus of contemporary american english (coca). Available online at https://www.english-corpora.org/coca/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of NAACL*, pages 4171–4186, Minneapolis, Minnesota.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2014. Using structured events to predict stock price movement: An empirical investigation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1415–1425, Doha, Qatar. Association for Computational Linguistics.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *Twenty-fourth international joint conference on artificial intelligence*.

Yo Ehara. 2018. Building an English Vocabulary Knowledge Dataset of Japanese English-as-a-Second-Language Learners Using Crowdsourcing. In *Proc. of LREC*.

Yo Ehara. 2019. Uncertainty-Aware Personalized Readability Assessments for Second Language Learners. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1909–1916.

Yo Ehara. 2020a. Interpreting neural CWI classifiers' weights as vocabulary size. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 171–176, Seattle, WA, USA → Online. Association for Computational Linguistics.

Yo Ehara. 2020b. Neural rasch model: How do word embeddings adjust word difficulty? In *Computational Linguistics*, pages 88–96, Singapore. Springer Singapore.

Yo Ehara, Yukino Baba, Masao Utiyama, and Eiichiro Sumita. 2016. Assessing Translation Ability through Vocabulary Ability Assessment. In *Proc. of IJCAI*.

Yo Ehara, Yusuke Miyao, Hidekazu Oiwa, Issei Sato, and Hiroshi Nakagawa. 2014. Formalizing Word Sampling for Vocabulary Prediction as Graph-based Active Learning. In *Proc. of EMNLP*, pages 1374–1384.

Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2012. Mining Words in the Minds of Second Language Learners: Learner-Specific Word Difficulty. In *Proceedings of COLING 2012*, pages 799–814, Mumbai, India. The COLING 2012 Organizing Committee.

Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2018. Mining Words in the Minds of Second Language Learners for Learner-specific Word Difficulty. *Journal of Information Processing*, 26:267–275.

Yo Ehara, Nobuyuki Shimizu, Takashi Ninomiya, and Hiroshi Nakagawa. 2013. Personalized Reading Support for Second-language Web Documents. *ACM Trans. Intell. Syst. Technol.*, 4(2):31:1–31:19.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Robert Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proc. of ICLR*.

Batia Laufer and Geke C. Ravenhorst-Kalovski. 2010. Lexical Threshold Revisited: Lexical Text Coverage, Learners' Vocabulary Size and Reading Comprehension. *Reading in a Foreign Language*, 22(1):15–30.

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.

John Lee and Chak Yan Yeung. 2019. Personalized Substitution Ranking for Lexical Simplification. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 258–267, Tokyo, Japan. Association for Computational Linguistics.

John SY Lee. 2021. An editable learner model for text recommendation for language learning. *ReCALL*, pages 1–15.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and Unsupervised Neural Approaches to Text Readability. *Computational Linguistics*, 47(1):141–179.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

I. Nation. 2006. How Large a Vocabulary is Needed For Reading and Listening? *Canadian Modern Language Review*, 63(1):59–82.

Gustavo Paetzold and Lucia Specia. 2016. Collecting and Exploring Everyday Language for Predicting Psycholinguistic Properties of Words. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1669–1679, Osaka, Japan. The COLING 2016 Organizing Committee.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, CINCINNATI UNIV OH.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Chak Yan Yeung and John Lee. 2018. Personalized Text Retrieval for Learners of Chinese as a Foreign Language. In *Proc. of COLING*, pages 3448–3455.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo H. Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. *arXiv:1804.09132 [cs]*. ArXiv: 1804.09132.