Text Retrieval for Language Learners: Graded Vocabulary vs. Open Learner Model

John S. Y. Lee, Chak Yan Yeung
Department of Linguistics and Translation
City University of Hong Kong
Hong Kong SAR, China
jsylee@cityu.edu.hk, cyyeung91@gmail.com

Abstract

A text retrieval system for language learning returns reading materials at the appropriate difficulty level for the user. The system typically maintains a learner model on the user's vocabulary knowledge, and identifies texts that best fit the model. As the user's language proficiency increases, model updates are necessary to retrieve texts with the corresponding lexical complexity. We investigate an open learner model that allows user modification of its content, and evaluate its effectiveness with respect to the amount of user update effort. We compare this model with the graded approach, in which the system returns texts at the optimal grade. When the user makes at least half of the expected updates to the open learner model, simulation results show that it outperforms the graded approach in retrieving texts that fit user preference for new-word density.

1 Introduction

Since language learning requires extensive extracurricular reading, learners can benefit from a text retrieval system that helps them identify suitable reading materials from a pool of candidate texts (Brown and Eskenazi, 2004; Miltsakaki, 2009; Lee, 2021). The suitability of a text may depend on multiple factors, including the user's reading interests (Heilman et al., 2007b), and the degree of matching between its difficulty and the user's proficiency. This paper investigates the use of an open learner model (OLM) to predict the latter.

By giving users more control over their learning, OLMs have been shown to foster users' confidence and reflection on their progress (Bull and Kay, 2007). These benefits are especially important for long-term or life-long learning activities (Kay and Kummerfeld, 2019), such as foreign language learning. We evaluate the effectiveness of an editable OLM (Bull and Kay, 2010) — i.e., an OLM that allows the user not only to view but also to modify its content — for text retrieval for language learning. As their vocabulary expands, users can update the OLM so that the system continues to retrieve texts that are lexically challenging to them. Specifically, we address two research questions:

Text retrieval performance (Q1): How accurately can an OLM identify reading materials with the desired density of new vocabulary, as specified by the user?

User update effort (Q2): How frequently does the user need to edit the OLM in order to reap its benefits?

Most previous research in computer-assisted language learning measured text retrieval performance through holistic evaluation (Heilman et al., 2007a), or in terms of users' overall learning outcomes (Hsu et al., 2013). In answering Q1, we will directly evaluate the density of new vocabulary in the retrieved texts. In addressing Q2, we will further consider how retrieval performance is affected by users' vocabulary acquisition over time and the amount of user update effort.

The rest of the paper is organized as follows. After a review of previous work (Section 2), we define our text retrieval framework (Section 3). We then describe and motivate the user simulation (Section 4). Next, we present the OLM approach and the baseline graded approach (Section 5). Finally, we compare the performance of the OLM and the graded approach with respect to the amount of user update effort (Section 6).

2 Previous work

While text difficulty can be influenced by a variety of lexical, syntactic, semantic and discourse features, many text recommendation systems focus on vocabulary (Brown and Eskenazi, 2004; Hsu et al.,

Proceedings of Recent Advances in Natural Language Processing, pages 798–804
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_091

2013; Wu, 2016), likely due to the strong correlation between vocabulary difficulty and text difficulty (Heilman et al., 2007b; François and Fairon, 2012). Our study will similarly adopt vocabulary difficulty as the text retrieval criterion.

Various approaches have been proposed for matching language learners with reading materials at a suitable level of lexical difficulty. The graded approach, also known as the leveling approach, places users and documents on a common scale, such as school grades (Miltsakaki and Troutt, 2008; Collins-Thompson et al., 2011). The system performs automatic readability assessment on each document and labels it with a grade to reflect its difficulty. This approach may not be able to capture individual learning patterns, however, as users are pigeon-holed into pre-defined grades.

As an alternative, the adaptive approach identifies user traits and preferences, and then adjusts the pedagogical content in the system to optimize learning outcomes (Brusilovsky, 2012; Vandewaetere et al., 2011). In the context of text retrieval for language learning, the system typically maintains a learner model on the user's linguistic proficiency, and then returns texts that best fit the model. The learner model may be estimated through user updates (Lee, 2021); formal assessment such as cloze items (Heilman et al., 2010); complex word identification models trained on vocabulary self-assessment by the user (Yeung and Lee, 2018); time log and click history patterns (Hokamp et al., 2014); as well as dictionary and translation queries by the user (Wu, 2016), among other non-invasive methods.

One disadvantage of the graded approach is the "jump" in difficulty when promoting the user from one grade to the next. The adaptive approach can potentially provide more fine-grained adjustments by gradually raising the vocabulary difficulty in the retrieved texts. To the best of our knowledge, there has not been any direct, quantitative comparison between these two approaches during a period of vocabulary acquisition by the user. This paper aims to fill in this gap.

3 Text retrieval framework

After motivating the use of vocabulary difficulty as the retrieval criterion (Section 3.1), we describe its implementation in the learner model (Section 3.2) and its application in the retrieval model (Section 3.3).

3.1 Retrieval criterion

The ideal text should have an appropriate amount of new vocabulary, so that it stretches the reader's competence without hindering comprehension. We quantify vocabulary difficulty by **new-word density** (Holley, 1973) (NWD), i.e., the percentage of words in the text that are *new for the user*. This metric is more straightforward to interpret and more transparent than grades, since users can easily examine the basis of retrieval results.

The system aims to return texts at a **Target NWD** that is specified by the user. The user thus has the freedom to set a relatively high Target NWD, for example, to maximize vocabulary acquisition, or set a relatively low one for leisure reading without dictionary look-ups.

3.2 Learner model

A language learner knows only a limited number of words in the foreign language. For each user u, we refer to this set of words as his or her *vocabulary set*, denoted as $voc(u) = \{w_1, \ldots, w_n\}$. Although nuances in lexical knowledge may be more precisely expressed with a real-number score (Yimam et al., 2018) or on a Likert scale (Ehara et al., 2012; Shardlow et al., 2021), we opted for the simpler known/unknown distinction to enable an intuitive interpretation of the NWD metric.

Since the system does not know the ground-truth vocabulary set $voc(u)$, the learner model needs to make an estimation $voc(\hat{u})$ for each user u. It can be effective to use automatic methods to re-estimate the vocabulary set as the user acquires new vocabulary (Section 2). However, we choose to base our evaluation on manual edits to an open learner model (OLM). This methodology has the advantage of being agnostic to the update algorithm, which may include any combination of manual and automatic methods, and may vary from one text retrieval system to another. Our results will therefore not be tied to any particular algorithm, but rather measure text retrieval performance with respect to varying amounts of valid updates (Section 5.1).

3.3 Retrieval model

The NWD of a document varies according to the user's vocabulary set. Formally, given a document d with D words, say $d = [w_1, \ldots, w_D]$, its **Actual**

NWD for user u is:

$$\frac{1}{D}\sum_{i=1}^{D} new_u(w_i) \quad (1)$$

where $new_u(w) = 0$ if the word $w \in voc(u)$, and $new_u(w) = 1$ otherwise.

Again, since the system has no access to the ground truth $voc(u)$, it must use $voc(\hat{u})$ to compute an **Estimated NWD**. The retrieval model returns the text whose Estimated NWD is closest to and not exceeding the Target NWD as specified by the user (Section 3.1).

4 Methodology

After motivating the advantages of using a simulation to compare the open learner model (OLM) and the graded vocabulary approach (Section 4.1), we give details on the simulation set-up (Section 4.2) and implementation details (Section 4.3) and define the evaluation metrics (Section 4.4).

4.1 Human subjects vs. user simulation

In the context of this study, text retrieval performance can be influenced by two variables: the Target NWD (Section 3.1) and the frequency of user update to the learner model (Section 3.2). It is therefore helpful to consider multiple configurations of these two variables.

There are a number of trade-offs between a user study and a user simulation. In the former, the subjects would need to perform text searches over a sufficiently long period of time to allow for substantial vocabulary acquisition. Throughout this period, they would need to read the retrieved texts and exhaustively annotate the unknown words therein, while experimenting with various update frequencies. This design has the advantage of providing authentic human data on vocabulary acquisition. However, it would introduce confounding factors such as differences among the subjects' proficiency levels, ability to work with the user interface, and diligence in updating the learner model. These factors are difficult to control for but can significantly influence the experimental results.

A user simulation can facilitate a more rigorous comparison by keeping these factors constant. It can also cheaply evaluate a large number of text searches, with no constraint on the length of the experimental period. The main disadvantage is that the users' vocabulary acquisition would need to be prescribed rather than empirically observed. This issue can be partially mitigated by consulting vocabulary lists, such as the widely used *Hanyu Shuiping Kaoshi* (HSK), which were crafted by experts with support from empirical data to reflect typical language learners (Hanban, 2014).

Given our research goals, we feel that the overall advantages of a simulation outweigh its disadvantages. Our simulation will be able to evaluate over 6K recommended documents in various experimental settings, a set of data points that is an order of magnitude larger than what we would have been able to gather from human subjects.

4.2 Simulation set-up

We simulated a user who searches for extracurricular reading materials for learning Chinese as a foreign language. We ran the simulation three times, with the Target NWD parameter set to $m\%$ NWD, for $m = \{20, 30, 40\}$.

Text retrieval. At times $i = 1, \ldots, k$, the user performs a text search to obtain documents whose Estimated NWD is closest to and not exceeding $m\%$ (Section 3.3). The user reads the top-ranked document that he or she has not yet read, and updates the OLM while reading (Section 5.1). Let d_i represent the document read by the user at the i^{th} search.

Vocabulary acquisition. Between two consecutive searches, the user learns a number of new words. Let u_i represent the user at the i^{th} search, and let W_i represent the set of new words learned between the i^{th} and $(i+1)^{\text{th}}$ searches. The user's vocabulary set expands during this period as follows:

$$voc(u_{i+1}) \leftarrow voc(u_i) \cup W_i \quad (2)$$

4.3 Simulation implementation

Let V_l denote the accumulative set of words in the HSK graded vocabulary lists up to level l, for $l = 1, \ldots, 6$. We set $voc(u_0) = V_5$ and $voc(u_k) = V_6$. This means that the user initially knows all the words listed up to HSK level 5, and then learns the words at level 6 during the simulation period.

We set a uniform learning rate at $|W_i| = 6$, meaning that six words are learned between two searches. The acquisition order is in reverse of word frequency in Chinese Wikipedia.

4.4 Evaluation metrics

We use two metrics to evaluate text retrieval performance:

NWD Error The difference between the Actual NWD of d_i and the Estimated NWD of d_i. This metric measures how much the estimated difficulty of the recommended document deviates from the ground truth. Recall that Estimated NWD is computed according to the estimated vocabulary set $voc(\hat{u}_i)$, while Actual NWD is based on $voc(u_i)$. These two figures differ whenever new words learned by the user appear in d_i but have not been updated in the learner model.

NWD Gap The difference between the Actual NWD of d_i and the Target NWD (Section 3.1). This metric expresses the discrepancy between the actual difficulty of the recommended document and the difficulty requested by the user.

5 Approach

As users learn new words, a document's new-word density (NWD) decreases. Periodic updates to the vocabulary set and re-estimation of the NWD of candidate documents are therefore necessary to ensure that the retrieved texts remain adequately challenging. We compare two approaches for this task.

5.1 Open learner model (OLM) approach

In the OLM approach, users are expected to manually update the vocabulary set described in Section 3.2. Hence, user update frequency crucially affects retrieval performance. If a user reports all newly acquired word to the OLM before using the text retrieval system, there would be no NWD Error. In practice, the user will likely update these words only after he or she encounters them when reading a recommended document d_i. At the time of search, the newly learned words would remain outside the vocabulary set and contribute to the NWD Error. We experimented with two update frequencies:

Full Update This frequency models the conscientious user who updates *all* words (that require update) when he or she reads the top-ranked documents d_i. Hence, following the i^{th} search, the vocabulary set is updated as follows:

$$voc(\hat{u}_{i+1}) \leftarrow voc(\hat{u}_i) \cup \Big(d_i \cap voc(u_i)\Big) \quad (3)$$

Occasional Update This models the more casual or conservative user who performs update on only half of the words in d_i that require update. In this more realistic scenario, newly learned words may remain excluded from the vocabulary set $voc(\hat{u}_i)$ even after the user has read them in multiple documents.

5.2 Graded Approach

Akin to a graded reader, the graded approach relies on the user to choose his or her grade, and assigns the vocabulary list corresponding to that grade as the vocabulary set. To create a strong baseline, we assume that the user always chooses the optimal grade. More formally, at time i, the graded approach uses the vocabulary list V_l that achieves the highest F-measure for the ground-truth vocabulary set $voc(u_i)$. In our case, then, the graded approach uses V_5 in the first half of simulation, and then switches to V_6 at the optimal time, when the user's lexical knowledge becomes closer to level 6.

This set-up gives the graded approach several advantages over the OLM. The graded approach not only selects the optimal grade, but also "knows" the words that the user will be learning in the simulation (namely, those in V_6). The OLM, in contrast, has no access to V_6 and relies only on user updates. Our simulation result will gauge the amount of user update necessary to reap the benefits of OLM over the graded approach.

6 Results

We conducted the simulation with a database of 1923K Chinese Wikipedia entries and 29K short essays.[1] We performed automatic word segmentation on all documents with the Stanford CoreNLP parser (Manning et al., 2014). We now present a chronological analysis of the simulation (Section 6.1), and then examine the overall experimental results (Section 6.2).

6.1 Chronological analysis

Figure 1 plots the Actual NWD of the top-ranked documents d_i over the course of the simulation, with the Target NWD set to 20%.

OLM approach (Full Update). Throughout the simulation, the model retrieved documents whose Actual NWD was relatively close to the 20% target, with the NWD Gap never exceeding 2%. The

[1] The short essays were downloaded from the website duanmeiwen.com

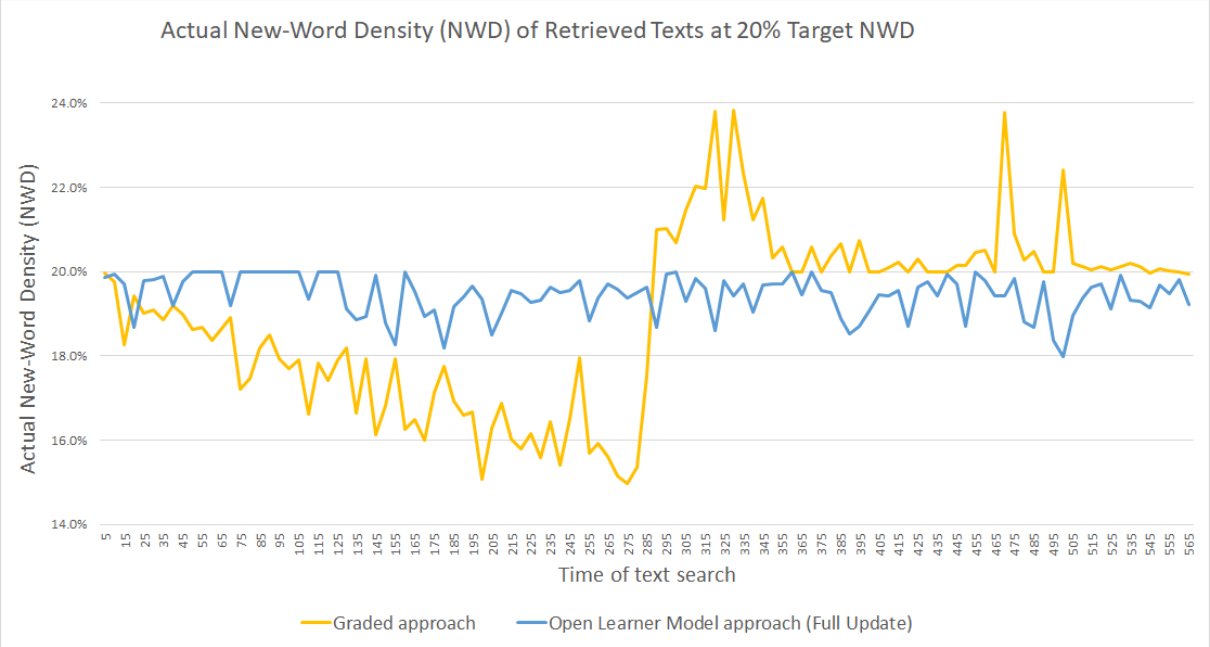

Figure 1: Actual new-word density (NWD) of the top-ranked document upon text retrieval at 20% Target NWD

Actual NWD was consistently below the target because the user would not update the status of newly learned words until he or she reads them in d_i.

Graded approach. In contrast, the NWD Gap for the graded approach reached a maximum of 5%. The Actual NWD of the recommended documents was initially close to the 20% target. With the vocabulary set kept constant at HSK level 5 during the first half of the simulation, the user's vocabulary acquisition led to a widening of the NWD Gap, up to 5% towards the middle of the simulation. At this point, with the promotion of the user to level 6, the NWD spiked to as high as 24.0%. The Actual NWD then gradually converged back to the 20% target towards the end of the simulation.

Approach	**Target NWD**	**NWD Gap**	**NWD Error**
Graded	20%	1.69%	1.35%
OLM (Occasional)		0.84%	0.84%
OLM (Full)		**0.55%**	**0.55%**
Graded	30%	2.27%	2.24%
OLM (Occasional)		1.26%	1.26%
OLM (Full)		**0.84%**	**0.84%**
Graded	40%	2.61%	2.59%
OLM (Occasional)		1.44%	1.44%
OLM (Full)		**1.07%**	**1.07%**

Table 1: New-word density (NWD) Gap and NWD Error of the top-ranked documents returned by the graded approach and the open learner model (OLM)

6.2 Overall results

Table 1 reports the average NWD Gap and NWD Error over the entire simulation. The OLM outperformed the graded approach at both update frequencies (full and occasional) and at all three targets (20%, 30% , 40%). We first analyze the results at 20% Target NWD, and then examine the effects of higher targets.

Target NWD at 20%. The OLM achieved the smallest NWD Gap with Full Update, at only 0.55% below the target. Aided by incremental adjustment to the vocabulary set, it more accurately re-estimated the user's current vocabulary competence, which in turn led to better NWD estimation. Occasional Update made the OLM more prone to over-estimate the difficulty of the documents, and hence produced a larger gap (0.84%). The graded approach incurred the largest NWD Gap, with an average of 1.69%. The gap was largest when the user was half-way between levels 5 and 6, since it was forced to choose one of the two and could not offer a middle ground.

In terms of NWD Error, the OLM also outperformed the graded approach at both update frequencies. The OLM was able to retrieve documents whose NWD more closely fits the user's target.

Effects of higher Target NWD. At higher NWDs, a larger pool of candidate documents can

fit the search criteria. In general, a document with more difficult words exposes the OLM to more chances of failing to recognize the user has learned those words. As a result, the higher the Target NWD, the larger the NWD Gap, i.e., the farther the recommended documents fell short of the target. For the OLM with Full Update, NWD Gap increased from 0.55% (Target=20%) to 0.84% (Target=30%) and 1.07% (Target=40%). A similar increase can be observed in the Occasional Update setting. These experimental results suggest that text retrieval is more challenging when the user requests documents with more advanced vocabulary.

7 Conclusions

Automatic text retrieval supports language learners in self-directed reading and independent learning. A major challenge in this task is to match learners with different capabilities to texts with appropriate vocabulary complexity.

We have evaluated an open learner model (OLM) that allows users to update their individual progress in vocabulary acquisition. We compared this model to the graded approach, where the system recommends texts to users at the optimal grade. We conducted a simulation of a learner of Chinese as a foreign language who uses the retrieval system during a period of vocabulary acquisition. Results show that the OLM outperforms the graded approach in retrieving texts at a range of target NWDs. When the user makes at least half of the expected updates, the OLM's fine-grained, incremental adjustment yields superior retrieval performance.

We believe these results can help inform the design of text retrieval systems for language learners. In future work, we intend to further improve retrieval quality by extending the OLM beyond vocabulary to other dimensions of text difficulty, such as syntactic and semantic complexity.

Acknowledgments

We gratefully acknowledge support of the Innovation and Technology Fund (Ref: ITS/389/17) of the Innovation and Technology Commission, the Government of the Hong Kong Special Administrative Region; and of the Hong Kong Institute for Data Science (Project #9360163) at City University of Hong Kong.

References

Jonathan Brown and Maxine Eskenazi. 2004. Retrieval of authentic documents for reader-specific lexical practice. In *Proc. InSTIL/ICALL Symposium*, Venice, Italy.

Peter Brusilovsky. 2012. Adaptive hypermedia for education and training. In *Adaptive technologies for training and education*, pages 46–66. Cambridge University Press, New York, NY, USA.

Susan Bull and Judy Kay. 2007. Student models that invite the learner in: The smili open learner modelling framework. *International Journal of Artificial Intelligence in Education*, 17(2):89–120.

Susan Bull and Judy Kay. 2010. Open Learner Models. In *Advances in Intelligent Tutoring Systems. Studies in Computational Intelligence, vol 308*, Berlin, Heidelberg. Springer.

Kevyn Collins-Thompson, Paul N. Bennett, Ryen W. White, Sebastian de la Chica, and David Sontag. 2011. Personalizing Web Search Results by Reading Level. In *Proc. 20th ACM International Conference on Information and Knowledge Management (CIKM)*.

Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2012. Mining words in the minds of second language learners: learner-specific word difficulty. In *Proc. International Conference on Computational Linguistics (COLING)*.

Thomas François and Cédric Fairon. 2012. An "AI Readability" Formula for French as a Foreign Language. In *Proc. EMNLP-CONLL*.

Hanban. 2014. *International Curriculum for Chinese Language and Education*. Beijing Language and Culture University Press, Beijing, China.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007a. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 460–467.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2010. Personalization of reading passages improves vocabulary acquisition. *International Journal of Artificial Intelligence in Education*, 20:73–98.

Michael Heilman, Alan Juffs, and Maxine Eskenazi. 2007b. Choosing reading passages for vocabulary learning by topic to increase intrinsic motivation. In *Proc. International Conference on Artificial Intelligence in Education*.

Chris Hokamp, Rada Mihalcea, and Peter Schuelke. 2014. Modeling Language Proficiency Using Implicit Feedback. In *Proc. 9th International Conference on Language Resources and Evaluation (LREC)*.

Freda M. Holley. 1973. A Study of Vocabulary Learning in Context: The Effect of New-Word Density in German Reading Materials. *Foreign Language Annals*, 6(3):339–347.

Ching-Kun Hsu, Gwo-Jen Hwang, and Chih-Kai Chang. 2013. A personalized recommendation-based mobile learning approach to improving the reading performance of EFL students. *Computers and Education*, 63:327–336.

Judy Kay and Bob Kummerfeld. 2019. From Data to Personal User Models for Life-long, Life-wide Learners. *British Journal of Educational Technology*.

John S. Y. Lee. 2021. An editable learner model for text recommendation for language learning. *ReCALL*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proc. ACL System Demonstrations*, pages 55–60.

Eleni Miltsakaki. 2009. Matching readers' preferences and reading skills with appropriate web texts. In *Proc. EACL Demonstrations Session*.

Eleni Miltsakaki and Audrey Troutt. 2008. Real time web text classification and analysis of reading difficulty. In *Proc. Third Workshop on Innovative Use of NLP for Building Educational Applications*.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 Task 1: Lexical Complexity Prediction. In *Proc. 15th International Workshop on Semantic Evaluation (SemEval)*.

Mieke Vandewaetere, Piet Desmet, and Geraldine Clarebout. 2011. The contribution of learner characteristics in the development of computer-based adaptive learning environments. *Computers in Human Behaviour*, 27:118–130.

Ting-Ting Wu. 2016. A learning log analysis of an English-reading e-book system combined with a guidance mechanism. *Interactive Learning Environments*, 24(8):1938–1956.

Chak Yan Yeung and John Lee. 2018. A Personalized Text Retrieval System for Learners of Chinese as a Foreign Language. In *Proc. International Conference on Computational Linguistics (COLING)*.

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. In *Proc. 13th Workshop on Innovative Use of NLP for Building Educational Applications*.