SATLab at BEA 2026 Shared Task 1: Predicting the Difficulty of English Words for Three L1 Learners Using Primarily Psycholinguistic Features

Yves Bestgen
Statistical Analysis of Text Laboratory (SATLab)
Faculté de psychologie, de logopédie, de sexologie et des sciences de la famille
Université catholique de Louvain
Place Cardinal Mercier, 10 1348 Louvain-la-Neuve, Belgium
yves.bestgen@uclouvain.be

Abstract

This paper presents SATLab's participation in the BEA 2026 shared task on predicting the difficulty of English words for L2 learners. The proposed system uses features mainly derived from word frequency lists, lexical norms, and psychometric data, which are input into a gradient boosting decision tree model. It outperformed the Baseline system but performed significantly worse than the top-performing teams. Feature contributions to model performance are analysed using gain scores and Spearman rank correlations, and a brief analysis of the most significant errors is provided.

1 Introduction

This paper presents SATLab's participation in the British Council's Shared Task at BEA 2026, *Vocabulary Difficulty Prediction for English Learners*, whose aim is to foster the development of regression models for predicting the difficulty of English words for learners of three first languages: Spanish, German, and Mandarin. Such models could, for example, be used to tailor content for learners at different proficiency levels.

To build high-performing models for this task, as in many other areas of NLP, the use of precomputed embeddings and LLMs has proven most effective (Skidmore et al., 2025; Yaneva et al., 2024). However, it is worth considering whether other features, such as word frequency lists, lexical norms, and psychometric data, can rival the simplest of these approaches and contribute to greater effectiveness. This seems particularly justified because Shardlow et al. (2021) observed, in a shared task on assessing word difficulty, little difference in effectiveness between deep learning approaches and feature-based approaches such as the system developped by Mosquera (2021). Previously, Culligan (2015) found that frequencies from very large corpora provide the best estimate of word difficulty (see also Settles et al. (2020)).

For several years, SATLab has specialised in using these indices to solve complex tasks such as predicting eye saccades during reading or identifying offensive content and hate speech in languages with few linguistic resources (Bestgen, 2021a,b,c). The vocabulary difficulty prediction task is particularly relevant to this objective due to the sustained interest it has attracted in linguistics (Schmitt and Schmitt, 2020) and the availability of 'fine-grained' resources based on extensive linguistic knowledge. The system proposed by SATLab is based on these features, which are fed into a gradient boosting decision tree model. One version of the system uses only these psycholinguistic features (*Nonly* version). The other version adds them to the difficulty score predicted by the organisers' Baseline system, which uses pre-computed embeddings (Felice and Skidmore, 2026; Skidmore et al., 2025).

The next section presents the main characteristics of the shared task. The features, the most important part of the proposal, are then described in detail. Finally, the results of the challenge are reported, along with an analysis to determine which features are the most useful.

2 Data and Task

The data for this challenge is taken from the *Extended Knowledge-based Vocabulary Lists Dataset for NLP* (Schmitt et al., 2021; Skidmore et al., 2025). It consists of 7,516 words, for which the difficulty level for English learners was estimated using a translation task involving over 100,000 learners from three L1 backgrounds. For each target lemma, the test item included the first letter, the length and the part of speech of the target word in English, one or more translations of the word into the L1, and a sentence in the L1 featuring the word in the relevant context. Based on responses in each L1, a difficulty level was estimated using a psychometric procedure (see Schmitt et al. (2024)

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 985–991

for further details).

The material was divided into three sets: the Learning set ($N = 6,091$), the Development set ($N = 677$), and the Test set ($N = 748$). The task was offered in two tracks. For the closed track, LLMs were not permitted to train the system or combine several L1s. All methods were permitted for the open track. The performance metric is the Root Mean Square Error (RMSE). Each team was allowed to submit three runs for each track and each L1.

The organisers also provided, as the baseline, the system they presented at BEA 2025 (Skidmore et al., 2025), which uses a fine-tuned transformer based architecture. This system was used to predict the Development and Test sets from the Learning set of the shared task materials.

3 Feature Sets

3.1 Features Derived Only from the Items

- Simple statistics: the character length of the target word (TargetL), the source word in L1 (SourceL) and the context (ContextL), the POS tag of the target (PosTag), the first letter of the target (FirstLetter) and the number of proposed L1 words (NbrWordL1).
- Sim12: The orthographic (Levenshtein distance) and phonetic (Soundex algorithm[1]) similarity were calculated to estimate the degree of cross-linguistic similarity between the English target word and the L1 source word (Otwinowska and Szewczyk, 2019). These measures are calculated only for German and Spanish.

3.2 Features Extracted from Generic Resources

- FBfr: Herdagdelen and Marelli (2017) Facebook frequency lists for American English and British English based on approximately one billion tokens for each variety. These lists were extracted from publicly available English posts collected between November 2014 and January 2015.
- TWfr: Herdagdelen and Marelli (2017) Rovereto Twitter Corpus frequency on more than one billion tokens from 75 million tweets, collected between December 2010 and July 2011.
- HALFr: Lund and Burgess (1996) Hyperspace Analogue to Language (HAL) frequency norms for approximately 40,000 words provided by (Balota et al., 2007).
- GoogleFr: The Google Ngram frequency word list[2].
- USENETfr: Shaoul and Chris (2006) USENET Orthographic Frequencies derived from a corpus of more than seven billion words from posts collected between October 2005 and August 2006.
- BNCfr: The frequency of orthographic forms in the 100-million-word British National Corpus. Two frequencies were calculated: the total frequency of the orthographic form and the frequency taking into account the PosTag.
- BasicV: Graham et al. (1993) list of 850 words most frequently used by children in their writing[3]. These words are classified into 5 grades according to their degree of difficulty for children, with the most difficult words to spell highlighted.
- Glasgow: Scott et al. (2019) lexical norms for age of acquisition, arousal, concreteness, dominance, familiarity, valence, gender association, imageability and semantic size. These norms include 5,553 English words evaluated on average by 33 native English speakers.
- ELP: Balota et al. (2007) lexical characteristics and behavioural measures from lexical decision and naming tasks for more than 40,000 words as available in the English Lexicon Project.
- Latent: Knoph et al. (2024) latent lexical dimensions of Frequency, Complexity, Proximity, Polysemy, and Diversity for 2,060 high-frequency words from the General Service List, 1,051 general academic words from the Academic Word List, and 3,413 domain-specific words from the Academic Vocabulary List.

It may seem surprising to use so many frequency-related features that are more or less strongly correlated. However, this approach has proved particularly effective in the Cognitive Modeling and Computational Linguistics (CMCL) Shared Task on Eye-Tracking Data Prediction (Bestgen, 2021a), enabling the system to outperform all deep learning-based systems participating in the challenge.

[1] https://www.archives.gov/research/census/soundex

[2] https://github.com/hackerb9/gwordlist

[3] https://www.readingrockets.org/topics/writing/articles/basic-spelling-vocabulary-list

3.3 Features Extracted from L2 Language Learner Resources

- Oxford: The Oxford University Press list of the 3,000 most essential words for English language learners[4], classified according to their level of difficulty based on CEFR levels A1 to B2.
- L2fr: The frequency of orthographic forms in three corpora of learners of English as a foreign language: the Ten-thousand English Compositions of Chinese Learners Corpus (TECCL), the International Corpus Network of Asian Learners of English (ICNALE) and the First Certificate in English examination scripts (FCE).
- LexCom: Maddela and Xu (2018) lexicon of 15,000 frequent English words annotated for complexity by eleven fluent English speakers of different native languages. These words are drawn from the most frequent words in the Google Ngram Corpus.
- BKL2: Brysbaert et al. (2020) database identifying the English words most familiar to non-native speakers, compiled through a large-scale crowdsourcing study involving 62,000 words.

4 System

LightGBM free software (Ke et al., 2017), a well-known implementation of the gradient boosting decision tree approach, was used to build the regression model. The various parameters were optimised using Optuna. When the feature derived from the Baseline was not used (*Nonly* version), nine-fold cross-validation was used on the development and training sets to select the best model and determine the number of iterations (early stopping rule). Nine folds were used to maintain the ratio between the number of instances in the development and training sets.

When this feature was used, the criterion was performance on the Development set. In the *Dev* version of the system, the Development set was used as provided by the organisers, i.e. the 677 unique instances. In the *Boot* version, 100 bootstrap samples (with replacement) of 677 instances from this Development set were constructed at the start of the optimisation, and the performance of a set of parameters corresponded to the average performance across these 100 samples. As these two versions produced almost identical results during the test phase, the remainder of this paper will no longer refer to the *Boot* version, which is less conventional.

[4] https://www.oxfordlearnersdictionaries.com/wordlists/oxford3000-5000

5 Analysis and Results

5.1 Official Results

The official challenge results are presented in the organisers' summary paper (Felice and Skidmore, 2026). In both tracks and across the three L1s, the *Dev* version performed significantly better (mean RMSE = 1.019) than the Baseline (mean RMSE = 1.176), but it also performed significantly worse than the top-performing team (mean RMSE = 0.790). In the closed track, the *Nonly* version, which does not use pre-computed embeddings, was also more effective (mean RMSE = 1.136) than the Baseline (mean RMSE = 1.218), which relies on a fine-tuned transformer-based architecture. However, the *Nonly* version was considerably less effective than the *Dev* version, which uses the Baseline as an additional feature (mean RMSE = 1.025).

5.2 Feature Contribution to Model Performance

To assess the usefulness of the features for predicting lexical complexity, Table 1 presents the importance scores of the top ten features in each of the three models analysed and for the three languages. These are gain scores, indicating how much each feature contributes to model performance. The table shows the sum of the gains for all features derived from the same resource. To facilitate interpretation, the values correspond to the proportion of each score relative to the maximum score for that model. A value of 0.50 therefore indicates that this feature set is half as useful as the most useful feature set.

For the *Nonly* version, the most useful predictors are clearly Sim12, BKL2, and Oxford. The effectiveness of Sim12 likely stems from its ability to estimate cross-linguistic similarity between the English target word and the L1 source word (Otwinowska and Szewczyk, 2019). The inclusion of Oxford in this group is notable because, as mentioned above, it is a list of just 3,000 words divided into four categories. Although relatively old and not specific to L2 learners, ELP is also very useful. Regarding the frequency lists, those from Twitter and Facebook are by far the best. I must admit

Spanish		German		Mandarin	
Feature	Gain	Feature	Gain	Feature	Gain
		Closed *Nonly*			
Sim12	1.00	Sim12	1.00	BKL2	1.00
BKL2	0.60	BKL2	0.66	Oxford	0.80
Oxford	0.41	Latent	0.44	TWfr	0.40
ELP	0.27	ELP	0.26	ELP	0.38
Glasgow	0.24	Glasgow	0.17	FBfr	0.30
LexCom	0.16	FBfr	0.10	Latent	0.24
Latent	0.13	TWfr	0.10	Glasgow	0.18
TWfr	0.11	LexCom	0.09	LexCom	0.18
FBfr	0.07	TargetL	0.04	SourceL	0.15
BNCfr	0.05	BNCfr	0.04	BNCfr	0.10
		Closed *Dev*			
BL	1.00	BL	1.00	BKL2	1.00
Sim12	0.72	Sim12	0.68	BL	0.85
BKL2	0.42	BKL2	0.33	ELP	0.52
Glasgow	0.15	ELP	0.15	TWfr	0.39
ELP	0.12	Oxford	0.14	FBfr	0.28
Oxford	0.12	TWfr	0.10	Oxford	0.25
TWfr	0.12	Glasgow	0.09	Latent	0.25
LexCom	0.08	FBfr	0.08	BNCfr	0.24
FBfr	0.07	Latent	0.06	Glasgow	0.18
Latent	0.07	LexCom	0.05	LCFreq	0.17
		Open *Dev*			
BL	1.00	BL	1.00	BL	1.00
Sim12	0.58	Sim12	0.41	BKL2	0.54
BKL2	0.39	BKL2	0.26	TWfr	0.25
Glasgow	0.13	Glasgow	0.06	ELP	0.25
Oxford	0.10	ELP	0.05	Latent	0.21
TWfr	0.09	TWfr	0.04	LCFreq	0.17
ELP	0.08	Latent	0.04	Glasgow	0.17
LexCom	0.06	LexCom	0.03	LexCom	0.15
FBfr	0.06	FBfr	0.02	FBfr	0.15
Latent	0.04	LCFreq	0.01	BNCfr	0.10

Table 1: Importance scores (LightGBM gain)

that I do not know why the length of the L1 source (SourceL) is relatively important for Mandarin. At first glance, I thought it would be the length of the target text, as is the case with German.

When the Baseline is introduced into the model (*Dev* versions), it is almost always the best predictor. This is particularly true for the Open track, where it performs much better than the other predictors. In general, there is little difference between the two tracks, which is to be expected since the scores to be predicted are the same and the predictors are the same, except for the Baseline, which was not trained on the same data. Finally, the differences observed between the three L1s do not appear to be very significant.

5.3 Spearman rank correlation with the lexical complexity score

The importance scores presented in Table 1 are not independent. An index that is relatively effective at predicting lexical complexity may be masked by the presence of other, more effective and strongly correlated indices, such as frequency norms derived from different corpora.

To assess the significance of each feature set independently, Table 2 presents the Spearman rank correlation (ρ) between the features and the score to be predicted, for the ten most highly correlated feature sets. Only the best feature of each set is taken into account. Rank correlation is used because it is insensitive to monotonic transformations, as is the gradient boosting decision tree approach, and is therefore not affected by the choice of one monotonic transformation over another (e.g. logarithmic, square root, inverse, or even highly implausible ones).

In each of the three L1s, the correlations of the best predictors are very similar. However, these correlations are significantly higher in Mandarin than in the other two languages. This result may be linked the fact that most teams performed significantly better in the official competition results for Mandarin than for Spanish and German. The excellent quality of Oxford is particularly evident in Mandarin, where it is the most correlated feature with vocabulary complexity.

LexCom also warrants special attention. Table 1 shows that this predictor does not make a significant contribution to model performance. Correlation analysis, however, reveals that for all three L1s, it is the second-best predictor and very close to the best. Another notable observation is that Sim12 is clearly not among the most highly correlated predictors for Spanish, yet it is the most useful feature for the model in the *Nonly* version and the second most important in the *Dev* versions. This is likely because it provides information distinct from the other features, almost all of which are linked in some way to frequency in the language. Regarding frequency indices, we observe that frequencies derived from Google Ngrams are never among the top ten predictors, nor are those from Usenet and the BNC. Further analysis would be valuable to understand the reasons for these poor performances.

5.4 Analysis of the most significant errors

The previous section identifies the most useful features for each L1. However, this should not obscure the fact that the proposed system makes relatively significant errors. The aim of this section is to gain insights into potential patterns found in poorly estimated items. To this end, I have extracted, sep-

Spanish		German		Mandarin	
Feature	ρ	Feature	ρ	Feature	ρ
Glasgow	0.48	BKL2	0.49	Oxford	0.63
LexCom	0.48	LexCom	0.48	LexCom	0.61
BKL2	0.47	Glasgow	0.47	BKL2	0.61
Oxford	0.43	Sim12	0.46	TWfr	0.60
TWfr	0.42	TWfr	0.44	Latent	0.60
FBfr	0.41	Oxford	0.43	FBfr	0.59
Sim12	0.40	FBfr	0.43	L2fr	0.58
Latent	0.37	ELP	0.39	Glasgow	0.56
ELP	0.37	Latent	0.38	ELP	0.52
BasicV	0.32	BasicV	0.32	HALfr	0.50

Table 2: Spearman rank correlation (Rho)

arately for each L1, the 20 items with the highest absolute error values.

Firstly, there is very little overlap between the three L1s, as over 90% of the items are specific to a single L1. The most significant errors are almost equal to 5, whereas the scale of the judgements to be predicted ranges roughly from -6 to +5.

In 90% of these 60 items, the error is an underestimation of the item's difficulty. A detailed analysis of the most problematic items shows that the issue often arises because the target is a relatively rare usage of a word well-known in another grammatical category. Examples include 'clear' used as an adverb, which usually appears in idiomatic expressions; 'received' used as an adjective meaning 'generally accepted'; 'very' used as an emphatic adjective ('the very nature of'); and 'dinning' used as a noun.

Among the most serious errors is also an item in which the word from the source language does not appear verbatim in the Spanish sentence provided in context: *6842,lamp,noun,l___,faro,Se encendió un foco de forma automática.,-2.171975368*

6 Conclusion

The aim of this study was to propose features primarily derived from word frequency lists, lexical norms, and psychometric data that could compete with more complex approaches, such as those based on precomputed embeddings, in the British Council's Shared Task at BEA 2026. Although the proposed system outperformed the Baseline, it performed significantly worse than several other systems in predicting English vocabulary complexity. It therefore appears necessary to conclude that "the glass is rather half empty than half full". Several limitations of this study, which also represent avenues for future research, are outlined in the following section.

Limitations

As this study focuses on feature sets that do not fall within the scope of deep-learning approaches, and as I lack expertise in this field, I used the out-of-the-box baseline to improve efficiency and, above all, to determine whether the proposed features provide information that could be useful for these newer and more complex approaches. This is undoubtedly a weakness of the study. For example, there is no basis for claiming that the features which appear beneficial for the model would also be so if more effective LLMs had been used.

Another limitation of the study is that I approached the task as a challenge focused on achieving performance, rather than as a scientific problem to be understood. This is evident in the inclusion of features such as the Arousal dimension, whose conceptual relevance to the problem at hand appears very low, but which preliminary analyses indicated improved performance. The *Extended Knowledge-based Vocabulary Lists Dataset* is so rich that it would have warranted a study aimed more at understanding.

Despite this emphasis on performance, the proposed features (see Section 3) focus almost exclusively on the target word in English. The translation of the target word into the L1 and the L1 sentence featuring the word in the relevant context are not used. Therefore, I do not know whether they would have led to better performance. However, it should be noted that these features are considered by the Baseline, whose prediction is used as a feature by the *Dev* version of the system.

A fourth limitation I wish to highlight is that I did not use the ablation approach to assess the independent contribution of each set of features to overall performance. Ablation is undoubtedly a powerful method. There were two reasons why I did not proceed in this way. Firstly, this would have required restarting the parameter optimisation procedure, thus consuming significant computational resources. Secondly, the few trials conducted showed that the models optimised on the development set were not the best-performing ones on the test data. Consequently, interpretation of the results would have been overly uncertain.

Acknowledgments

The author wishes to thank the organizers of this shared task for putting together this valuable event. He is a Research Associate of the Fonds de la Recherche Scientifique - FNRS (Fédération Wallonie Bruxelles de Belgique).

References

David A. Balota, Melvin J. Yap, Keith A. Hutchison, Michael J. Cortese, Brett Kessler, Bjorn Loftis, James H. Neely, Douglas L. Nelson, Greg B. Simpson, and Rebecca Treiman. 2007. The English lexicon project. *Behavior Research Methods*, 39:445–459.

Yves Bestgen. 2021a. LAST at CMCL 2021 shared task: Predicting gaze data during reading with a gradient boosting decision tree approach. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 90–96, Online. Association for Computational Linguistics.

Yves Bestgen. 2021b. LAST at SemEval-2021 Task 1: improving multi-word complexity prediction using bigram association measures. In *Proceedings of SemEval-2021*.

Yves Bestgen. 2021c. A simple language-agnostic yet strong baseline system for hate speech and offensive content identification. In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation*, CEUR Workshop Proceedings. CEUR-WS.org.

Marc Brysbaert, Emmanuel Keuleers, and Paweł Mandera. 2020. Which words do English non-native speakers know? New supernational levels based on yes/no decision. *Second Language Research*, 36(4):395–417.

Brent Culligan. 2015. A comparison of three test formats to assess word difficulty. *Language Testing*, 32(4):503–520.

Mariano Felice and Lucy Skidmore. 2026. Findings of the BEA 2026 shared task on vocabulary difficulty prediction for English learners. In *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, San Diego, California. Association for Computational Linguistics.

Steve Graham, Karen R Harris, and Connie Loynachan. 1993. The basic spelling vocabulary list. *The Journal of Educational Research*, 86(6):363–368.

Amac Herdagdelen and Marco Marelli. 2017. Social media and language processing: How Facebook and Twitter provide the best frequency estimates for studying word recognition. *Cognitive Science*, 41:976–995.

Shin'ichiro Ishikawa. 2023. *The ICNALE Guide: An Introduction to a Learner Corpus Study on Asian Learners' L2 English*. Routledge.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3146–3154. Curran Associates, Inc.

Rebecca E. Knoph, Joshua F. Lawrence, and David J. Francis. 2024. The dimensionality of lexical features in general, academic, and disciplinary vocabulary. *Scientific Studies of Reading*, 28(4):319–342.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208.

Mounica Maddela and Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760, Brussels, Belgium. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Agnieszka Otwinowska and Jakub M. Szewczyk. 2019. The more similar the better? Factors in learning cognates, false cognates and non-cognate words. *International Journal of Bilingual Education and Bilingualism*, 22(8):974–991.

Norbert Schmitt, Karen Dunn, Barry O'Sullivan, Laurence Anthony, and Benjamin Kremmel. 2021. Introducing Knowledge-based Vocabulary Lists (KVL). *TESOL Journal*, 12(4):e622.

Norbert Schmitt, Karen Dunn, Barry O'Sullivan, Laurence Anthony, and Benjamin Kremmel. 2024. *Knowledge-based Vocabulary Lists*. University of Toronto Press.

Norbert Schmitt and Diane Schmitt. 2020. *Vocabulary in Language Teaching*. Cambridge University Press.

Graham G. Scott, Anne Keitel, Marc Becirspahic, Bo Yao, and Sara C. Sereno. 2019. The Glasgow norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, 51:1258–1270.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

Cyrus Shaoul and Westbury Chris. 2006. USENET orthographic frequencies for 111,627 English words.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Lucy Skidmore, Mariano Felice, and Karen Dunn. 2025. Transformer architectures for vocabulary test item difficulty prediction. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 160–174, Vienna, Austria. Association for Computational Linguistics.

Victoria Yaneva, Kai North, Peter Baldwin, Le An Ha, Saed Rezayi, Yiyun Zhou, Sagnik Ray Choudhury, Polina Harik, and Brian Clauser. 2024. Findings from the first shared task on automated prediction of difficulty and response time for multiple-choice questions. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 470–482, Mexico City, Mexico. Association for Computational Linguistics.

Helen Yannakoudakis and Ted Briscoe. 2012. Modeling coherence in ESOL learner texts. In *Proceedings of the 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, pages 33–43.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189.