Chen, W. et al. (Eds.) (2017). Proceedings of the 25th International Conference on Computers in Education.
New Zealand: Asia-Pacific Society for Computers in Education

Word Error Rate as a Listenability Index for Learners of English as a Foreign Language

Katsunori KOTANI[a*] & Takehiko YOSHIMI[b]
[a]*School of International Professional Development, Kansai Gaidai University, Japan*
[b]*Faculty of Science and Technology, Ryukoku University, Japan*
*kkotani@kansaigaidai.ac.jp

Abstract: In learning/teaching English as a foreign language, it is necessary to prepare listening materials that match learners' proficiency. Recent development of computer-assisted language learning/teaching solves a matching of proficiency by automatically measuring the ease of listening comprehension (listenability). Previously, an index for listenability was determined by learners' subjective judgment for listening comprehension. The present study proposed word error rate (WER) in transcription as an alternative listenability index. The experimental results demonstrated the reliability and validity of the WER as a listenability index.

Keywords: listenability, word error rate, learning material, English as a foreign language

1. Introduction

An advantage of computer-assisted language learning/teaching is the use of listening materials taken from the Internet, which results in a heavy burden on language teachers to check whether materials are appropriate for the proficiency of their learners so as to prevent decreases in learning motivation. A solution is to use an automatic measuring method for the ease of listening comprehension (listenability) (Kotani & Yoshimi 2016, Yoon et al. 2016). These previous studies determined an index for listenability by learners' subjective judgment for listening comprehension on a five-point Likert scale. Although this approach succeeded in measuring the listenability at the text level (Yoon et al. 2016) and at the sentence level (Kotani et al. 2016), it remained open for the possibility of measurement in more detail by measuring the listenability at the word level within context.

Therefore, the aim of the present study is to propose word error rate (WER) in transcription as another listenability index. This study also reports experimental results for the reliability and validity of WER as a listenability index, which was examined within the classical test theory (Brown 1996). The results showed that the WER was as reliable a listenability index as subjective judgment, and that the WER was a more valid listenability index than subjective judgment.

2. Compilation of Listenability Data

Listening materials were produced based on two texts distributed by the International Phonetic Association (1999), and the texts included all of the English phonemes. The voice actor (female, 35 years old, Canadian) read the texts aloud with an American accent at natural speech rate (approximately 180 words per minute (Rodero 2012)).

The listenability data were compiled from 50 learners of English as a foreign language at university (28 males, 22 females; mean age: 20.8 ± 1.3 years old who were compensated for their participation. All learners were asked to submit valid scores (10–990) from the Test of English for International Communication (TOEIC) in the current or previous year. In our sample, the mean TOEIC score was 607.7 ± 186.2.

WER data were derived based on evaluation results of learners' transcription by two university English teachers: (i) correct transcription, (ii) deletion, (iii) substitution, and (iv) addition, which were annotated using the UAM Corpus Tool (O'Donnell 2008). WER was calculated by dividing the number of transcription error tags (deletion, substitution, addition) by the total number of words in a reference

spoken sentence. In order to examine the inter-evaluator reliability, correlation analysis was performed between an evaluator's WER (WER-A, where the mean value was 0.59 ± 0.05) and the other's WER (WER-B, where the mean value was 0.56 ± 0.05), which showed strong correlation (r = .97). Subjective judgment data were derived from scores subjectively determined by learners on a five-point Likert scale (from 1: easy, 2: somewhat easy, 3: average, 4: somewhat difficult, or 5: difficult), following previous research (Kotani et al. 2016, Yoon et al. 2016).

3. Assessment of the WER as a Listenability Index

The reliability of the WER was examined using Cronbach's alpha (Cronbach 1970) defined in equation ($\alpha = \frac{k}{k-1}(1 - \sum_{i=1}^{k} \frac{S_i^2}{S_T^2})$, where α is the reliability coefficient, k is the number of items (here: sentences), S_i^2 is the variance associated with item i, and S_T^2 is the variance associated with the sum of all k-item values). Cronbach's alpha reliability coefficient ranges from 0 (absence of reliability) to 1 (absolute reliability), and empirical satisfaction is achieved with values above 0.8. The reliability coefficients for WER-A, WER-B, and the subjective judgments (WER-A: α = 0.97, WER-B: α = 0.97, and subjective judgment: α =0.89) outperformed the baseline.

The validity of WER was examined in terms of whether it reflected learner proficiency-dependent listenability. The dependence of listenability on a learner's proficiency refers to the situation in which listenability is higher for learners at a high proficiency level than for those at a low proficiency level. Along with TOEIC scores, mean WER and subjective judgment values of learners were calculated by dividing the sum of WER/subjective judgment values by the number of sentences (15 sentences).

The construct validity of WER was examined from the viewpoint of the distinctiveness between proficiency levels: beginner (TOEIC scores ranging from 295–450), intermediate (490–685), and advanced (730–900). Table 1 shows the mean (SD) values for WER-A, WER-B, and subjective judgment of these groups.

Table 1: WER and subjective judgment values by proficiency level.

	Beg. (n = 16)	Int. (n = 16)	Adv. (n = 18)
WER-A	0.69 (0.17)	0.64 (0.17)	0.40 (0.19)
WER-B	0.68 (0.17)	0.62 (0.17)	0.38 (0.19)
Subjective judgment	4.5 (0.7)	4.3 (0.7)	3.9 (0.9)

One-way ANOVA showed statistically significant differences ($p < 0.01$) between all three groups of learners for TOEIC scores (F (2, 47) = 235.4), WER-A (F (2, 47) = 39.7), WER-B (F (2, 47 = 41.6), and subjective judgments (F (2, 47 = 9.2). The results were further examined using Tukey's post-hoc comparison, which showed statistically significant ($p < 0.01$) in the WER and the subjective judgment between the beginner and advanced levels and between the intermediate and advanced levels, but not between the beginner and the intermediate levels.

The criterion-related validity of WER was examined from the viewpoint of the correlation with learners' TOEIC scores. TOEIC scores showed strong correlations with WER-A (r = –0.83) and WER-B (r = –0.84), but a weak correlation with subjective judgments (r = –0.51). According to the TOEIC technical manual (Chauncey Group International 1998), empirically valid correlation coefficients above 0.73 were found, as TOEIC scores were correlated with a valid English test (r = 0.73). The WER and subjective judgment were further examined in an asymptotic z-test with by using Fisher's z-transformation (Lee & Preacher 2013). Statistically significant differences ($p < 0.01$) were observed with WER-A (n = 50, z = –3.0), and with WER-B (n = 50, z = –3.2). That is, the WER was more valid than subjective judgment.

4. Conclusion

This study examined listenability indices by comparing subjective judgment of previous studies (Kotani et al. 2016, Yoon et al. 2016) with WER of transcription proposed by this study. The reliability and

validity of the WER, assessed using classical test theory, were confirmed as well as the subjective judgment. WER outperformed the subjective judgment in the criterion-related validity. This high criterion-related validity seems to be caused by the objective evaluation which excludes over/underestimation by learners.

The experimental results support the use of a listenability measurement method in a classroom, where a listenability measurement method is available as an education application. First, a teacher/learner picks up listening materials on the Internet. Second, listening materials are examined linguistically in order to extract linguistic features such as sentence length and speech rate. Third, linguistic features are input to a listenability measurement application, in addition to a learner feature of listening proficiency in terms of a TOEIC score. Then, the application calculates and demonstrates the listenability of the materials. According to the results, a teacher/learner chooses listening materials among three types of listenability: low, moderate, or high. Materials with high listenability should be chosen for extensive listening practice, and those with low listenability for intensive listening practice.

A remaining problem of this study is to seek another alternative listenability index that consists of both subjective and objective evaluations. Subjective judgment is plausible in that it directly demonstrates learners' listenability, which is difficult for WER to explain. The problem of subjective evaluation would be solved by combining with objective evaluation, that is, WER. Future study needs to assess the reliability and validity of a complex listenability.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. All remaining inadequacies are ours alone. This work was supported by JSPS KAKENHI Grant Numbers, 22300299, 15H02940.

References

Brown, J. D. (1996) *Testing in Language Programs*. New Jersey: Prentice Hall Regents.

Chall. J. S., & Dial, H. E. (1948) Predicting listener understanding and interest in newscasts. *Educational Research Bulletin*, 27(6), 141-153+168.

Chauncey Group International (1998) *TOEIC Technical Manual*. New Jersey: Chauncey Group International.

Cronbach, L. J. (1970) *Essentials of Psychological Testing*. New York: Harper & Row.

Fang, I. E. (1966) The "Easy listening formula." *Journal of Broadcasting & Electronic Media*, 11(1), 63-68.

International Phonetic Association (1999) *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge, UK.: Cambridge University Press.

Kiyokawa, H. (1990) A formula for predicting listenability: The listenability of English language materials 2. *Wayo Women's University Language and Literature*, 24, 57-74.

Kotani, K., & Yoshimi, T. (2016) Learner feature variation in measuring the listenability for learners of English as a foreign language. In Zou, D. et al. (Eds.) *Proceedings of the 1st International Workshop on Emerging Technologies for Language Learning* (unpaged).

Lee, I. A. & Preacher, K. J. (2013) *Calculation for the Test of the Difference between Two Dependent Correlations with One Variable in Common*. Available from http://quantpsy.org.

Messerklinger, J. (2006) Listenability. *Center for English Language Education Journal*, 14, 56-70.

O'Donnell, M. (2008) Demonstration of the UAM Corpus Tool for text and image annotation. In Arisoy, E. et al. (Eds.) *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies* (pp. 13-16).

Rodero, E. (2012) A comparative analysis of speech rate and perception in radio bulletins. *Text & Talk*, 32(3), 391-411.

Yoon, S.-Y., Cho, Y., & Napolitano, D. (2016) Spoken text difficulty estimation using linguistic features. In Tetreault, J. et al. (Eds.) *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 267-276).