Experimental analyses of the factors affecting the gradience in sentence difficulty judgments

Cesar Koirala[1] and Rebecca Y. Jee[2]

Abstract. Although a reader's text-level comprehension is affected by the comprehension of individual sentences in a text, little attention has been paid to the difficulty of sentences. This study investigates whether measures (features) of text difficulty affect the *gradience* observed in sentence difficulty judgments. We examine two traditional features (*sentence length* and *number of low-frequency words*) and six nontraditional features (counts of *clauses*, *dependent clauses*, *coordinate phrases*, *t-units*, *complex t-units*, and *Wh nominals*). Five English language instructors participated in a sentence difficulty behavioral experiment. They had to judge how difficult the sentences were on a scale of 1 to 4, where '1' means *very easy*, '2' means *easy*, '3' means *moderately difficult*, and '4' means *difficult*. The scale of 1-4 allowed the subjects to treat perceived difficulty as a relative point on the scale rather than as categorical values ('easy' and 'difficult'). It was found that both traditional and nontraditional measures of text difficulty correlate with participants' perceived difficulty suggesting that both types of features play roles in the perception of sentence difficulty. In addition, a tree-based readability model was implemented using the same sentences and features. The preliminary data suggests that the traditional features are more important while classifying new observations.

Keywords: machine learning, NLP-based features, sentence difficulty, text readability.

1. Voxy, New York, NY, USA; cesar@voxy.com

2. Voxy, New York, NY, USA; rebecca@voxy.com

How to cite this article: Koirala, C., & Jee, R. Y. (2015). Experimental analyses of the factors affecting the gradience in sentence difficulty judgments. In F. Helm, L. Bradley, M. Guarda, & S. Thouësny (Eds), *Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy* (pp. 324-329). Dublin: Research-publishing.net. http://dx.doi.org/10.14705/rpnet.2015.000353

1. Introduction

It is well known that learning is effective when learners are able to truly engage with the material presented to them. If a piece of material is too difficult, learners will not be able to understand or retain the information. For this reason, we find abundant readability research in both first language (L1) acquisition and second language (L2) acquisition. However, although it is well known that the comprehension of individual sentences in a text affects a reader's text-level comprehension, little attention has been paid to the difficulty of sentences (c.f. Scott, 2009).

As mentioned above, text readability has been the subject of several studies, and traditional formulas for computing text readability like Flesch Reading Ease (Flesch, 1948) and Flesch-Kincaid Grade Level (Kincaid, Fishburne, Rogers, & Chissom, 1975) date decades back. Most traditional formulas developed for assessing text difficulty are based on combinations of simple features like vocabulary *frequency* or sentence *length* (c.f Klare, 1984; DuBay, 2004). With the emergence of efficient Natural Language Processing (NLP) systems, new researchers have been able to use sophisticated features that need more time and resources to compute. Some of these studies have suggested that sophisticated NLP-based features combined with machine learning algorithms perform better than traditional readability formulas in predicting text difficulty (c.f Francois & Miltsakaki, 2012).

In this study, we are interested in the perceived difficulty of individual sentences, and not the entire text. Furthermore, we examine gradient difficulty of sentences (*very easy* to *difficult*) rather than categorical difficulty (*easy* and *difficult*). We ask the following questions in particular:

- How well do the measures of text difficulty predict sentence difficulty? We investigate the effects of two traditional features (*sentence length* and *number of low-frequency words*) and six nontraditional features (counts of *clauses*, *dependent clauses*, *coordinate phrases*, *t-units*, *complex t-units*, and *Wh nominals*) (Lu, 2010).

- Can these features explain the gradience in sentence difficulty perception?

2. Behavioral experiment

A behavioral experiment was devised to investigate whether the traditional and nontraditional features of text difficulty affect the gradience observed in sentence

difficulty judgments. It is shown that both types of features correlate with the subjects' perceived difficulty, suggesting that both types of features play roles in the perception of sentence difficulty. The components of the experiment are described in detail below.

2.1. Subjects

Five English language instructors were recruited from Voxy, an education technology startup in New York. The participants were unaware of the design and purpose of the experiment.

2.2. Stimuli

The stimuli consisted of 499 English sentences. A python script was used to pull the sentences randomly from the Voxy corpus. The corpus consisted of articles from authentic sources such as Oxford University Press, Bloomberg, the Associated Press, and the Financial Times. It also contained teaching materials prepared by the company's own publishing team. The stimuli covered a wide range of topics like sports, technology, entertainment, and politics, as exemplified below:

- Mexican soccer star Rafael Marquez may leave Barcelona.
- Amazon launched the $199 tablet last November.
- Is the world becoming more and more obsessed with covering celebrities?
- Republican attempts to amend the law will continue, he said, but outright repeal is no longer a possibility.

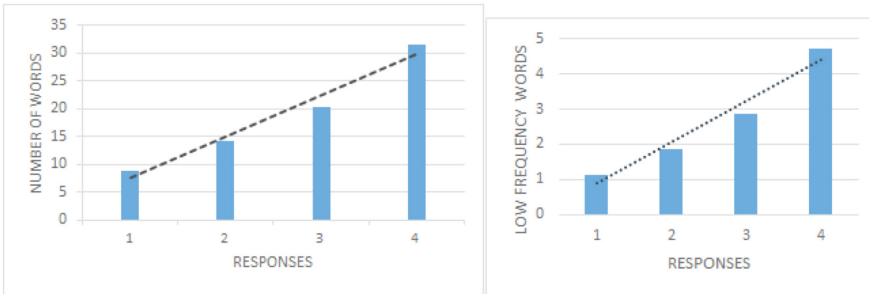
2.3. Procedure

The experiment was implemented using E-prime software (http://www.pstnet.com/eprime.cfm). Subjects saw English sentences on the computer screen and were asked to judge the difficulty of the sentences on a scale of 1 to 4, where '1' means *very easy*, '2' means *easy*, '3' means *moderately difficult*, and '4' means *difficult*. The subjects were instructed to make use of the whole scale as much as possible. In order to avoid the response bias (a general tendency to respond either yes or no), the scale of 1 to 4 was used rather than providing 'yes' or 'no' options to the subjects. This also allowed the subjects to treat perceived difficulty as a relative point on the scale rather than saying the sentences were exactly 'easy' or 'difficult.'

3. Results

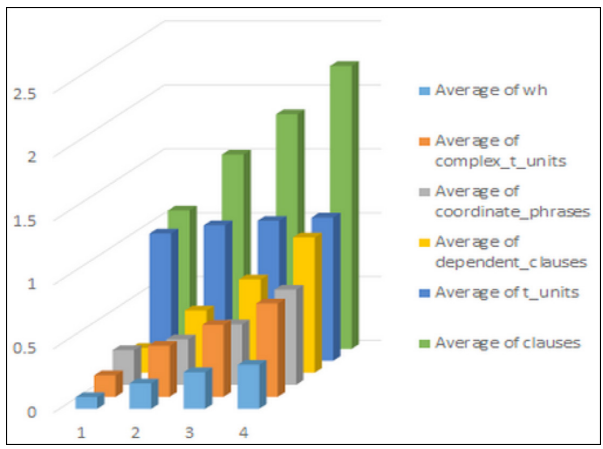
Figure 1 shows the average number of words and average number of low-frequency words per sentence for the values on the scale. Perceived difficulty increased with the increase in both traditional features.

Figure 1. Traditional features



Like traditional features, the perceived difficulty increased with the increase in the counts of nontraditional features, as shown in Figure 2.

Figure 2. Nontraditional features

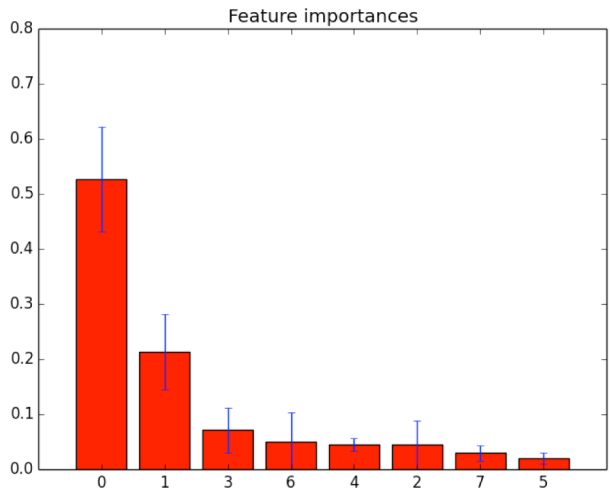


4. Feature importances

Using Scikit-learn (Pedregosa et al., 2011), a *random forest* classifier for sentence difficulty was implemented using the same sentences and features. The

random forest algorithm can be used to estimate the importance of variables in a classification task. The feature importances estimated by the random forest model for the data are plotted in Figure 3 below, where '0' is 'number of words,' '1' is 'number of low-frequency words,' and '2' to '7' represent nontraditional features.

Figure 3. Feature importances



Results show that traditional features outperform the nontraditional features as single predictors of difficulty.

5. Conclusions

We examined whether or not the measures of text difficulty affect the *gradience* observed in sentence difficulty judgments. Results of the behavioral experiment suggested that both traditional and nontraditional measures of text difficulty play roles in determining sentence difficulty. For all features, as the counts of features increased, perceived difficulty increased as well, exhibiting gradience of difficulty. We also implemented a *random forest* classifier for sentence difficulty using the same sentences and features. It was found that traditional features are more important as they are better as single predictors of difficulty. This could be due to overlaps between the two types of features. For instance, *counts of clauses* measures the length of production unit, and *sentence length* does the same, too. In the future, we plan to construct classifiers that use all 14 features in Lu (2010) to test whether these findings still hold.

It is important to note that the language instructors agreed on the perceived difficulty of sentences even without rubrics that help distinguish difficulty scales. As a follow-up experiment, we plan to conduct this same experiment on language learners and compare those findings with these current findings.

References

DuBay, W. H. (Eds.). (2004). *The principles of readability*. California: Impact Information.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*, 221-233. doi:10.1037/h0057532

Francois, T., & Miltsakaki, E. (2012). Do NLP and machine learning improve traditional readability formulas? *In Proceedings of the First Workshop on Predicting and Improving Text Readability for Target Reader Population, NAACL*.

Kincaid, J. P., Fishburne, R. P. Jr., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Research Branch Report*, 8-75. US: Naval Air Station, Memphis.

Klare, G. R. (1984). Readability. In P. D. Pearson, R. Barr, M. L. Kamil, & P. Mosenthal (Eds.), *Handbook of Reading Research* (pp. 681-744). New York, NY: Longman.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474-496. doi:10.1075/ijcl.15.4.02lu

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

Scott, C. M. (2009). A case for the sentence in reading comprehension. *Language, Speech, and Hearing Services in Schools, 40*, 184-191. doi:10.1044/0161-1461(2008/08-0042)

Published by Research-publishing.net, not-for-profit association
Dublin, Ireland; info@research-publishing.net

Critical CALL – Proceedings of the 2015 EUROCALL Conference, Padova, Italy
Edited by Francesca Helm, Linda Bradley, Marta Guarda, and Sylvie Thouësny

Typeset by Research-publishing.net
Fonts used are licensed under a SIL Open Font License

ISBN13: 978-1-908416-28-5 (Paperback - Print on demand, black and white)
Print on demand technology is a high-quality, innovative and ecological printing method; with which the book is never 'out of stock' or 'out of print'.

ISBN13: 978-1-908416-29-2 (Ebook, PDF, colour)
ISBN13: 978-1-908416-30-8 (Ebook, EPUB, colour)

Legal deposit, Ireland: The National Library of Ireland, The Library of Trinity College, The Library of the University of Limerick, The Library of Dublin City University, The Library of NUI Cork, The Library of NUI Maynooth, The Library of University College Dublin, The Library of NUI Galway.

Legal deposit, United Kingdom: The British Library.
British Library Cataloguing-in-Publication Data.
A cataloguing record for this book is available from the British Library.

Legal deposit, France: Bibliothèque Nationale de France - Dépôt légal: décembre 2015.