Longitudinal Analysis of Historical Texts' Readability

Adam Jatowt[1,2] and Katsumi Tanaka[1]

[1]Kyoto University
Yoshida-Honmachi, Sakyo-ku
606-8501 Kyoto, Japan

[2]Japan Science and Technology Agency
4-1-8, Honcho, Kawaguchi-shi, Saitama
332-0012 Tokyo, Japan

{adam, tanaka}@dl.kuis.kyoto-u.ac.jp

ABSTRACT

Digital libraries often contain historical documents of varying age. The degree to which users can understand their content depends much on their reading difficulty. In this poster paper we report the results of our studies on the readability of historical documents from the viewpoint of present users. We investigate the correlation between the outcomes of different readability measurements and publication dates of prose texts on the basis of two datasets, the *Victorian Women's Writers Project* and the *Corpus of Late Modern English Texts.*

Categories and Subject Descriptors

H.3.0 [**Information Storage and Retrieval**]: General

General Terms

Algorithms, Measurement, Experimentation

Keywords

readability, historical documents, language evolution

1. INTRODUCTION

Readability indices have been employed to measure complexity of various document types ranging from scientific papers, clinical reports, legal documents, textbooks to, more recently, web pages. However, little work has been devoted to approaching the readability of historical documents as seen from the viewpoint of present users. In recent years we have witnessed massive digitalization of historical text sources that are being preserved by libraries, museums and other institutions. This joint digitalization effort has resulted already in a large body of content that can be effectively measured, compared and described using computers.

Document readability is a key aspect of information accessibility and to ensure the access to information in historical documents their readability levels should be quantified. They could then serve as an indicator of the ease of reading of resources and should be made clear to users while browsing and searching old documents. Perhaps, also specialized tools such as ones using statistical translation models could be offered for readers to support understanding of old documents. Readability measure could be also used as a feature for automatically dating documents.

In this paper we investigate the correlation between readability of English prose and its age using two well-known datasets, *Victorian Women's Writers Project* (VWWPC) and the *Corpus of Late Modern English Texts* (CLMETEV) spanning different sub-periods within the last three centuries. Both the VWWPC and CLMETEV corpora have been created by professional linguists with the aim of being used for linguistic analysis.

JCDL'12, June 10–14, 2012, Washington, DC, USA.
ACM 978-1-4503-1154-0/12/06.

Sherman [2] first observed that the lengths of sentences decreased over time and written language become closer to spoken one. However, his work lacked quantitative analysis and was limited to the study of the sentence length effect on the readability. Danielson and Lasorsa [1] investigated whether prose can be dated using stylistic markers of sentence length, word length, rare punctuation, and word shortening or informality. They found positive correlation between the document age and these markers.

2. DATASETS AND METRICS

The first dataset, Victorian Women's Writers Project Corpus[1] (VWWPC) contains 200 original texts written by female writers from the 1820s to the 1920s. The corpus includes balanced samples of different genres such as poetry, novels, children's books, political pamphlets, religious tracts, histories etc.

The second dataset, *Corpus of Late Modern English Texts*[2] (CLMETEV), incorporates texts written by British native speakers with the constraint that no author contributes more than 200,000 words of text. Although the corpus texts vary in terms of genre, ranging from personal letters to literary fiction to scientific writing, its main part contains prose. The texts were written by both men and women of varying social class within the period of the 1720s to the 1920s. Table 1 shows the summary of the key statistics of both the datasets.

Table 1 Summary of key statistics of the datasets.

Dataset	#Resources	#Sentences	#Words	Period
VWWPC	200	319K	7mln	1830s-1920s
CLMETEV	176	665K	14,9mln	1720s-1920s

As readability measures we have used *Flesch Reading Ease* (FRE), *Coleman Liau Index* (CLI) which are common readability metrics and own readability measure based on the analysis of part-of-speech distributions. FRE is defined as 206.876 -1.015ASL – 84.6ASW where ASL is the average number of syllables per word and ASW is the average number of words per sentence. Higher FRE values indicate more readable text. On the other hand, CLI approximates the U.S. grade level thought to be necessary for comprehending any text. It is defined as $0.0588L – 0.296S – 15.8$ where L stands for the average number of letters per 100 words and S denotes the average number of sentences per 100 words. High values of CLI indicate difficult texts.

We have also created own measure of grammatical complexity, called *POSR*, based on the analysis of category sections of Encyclopedia Britannica[3] treated as a training set. We have taken

[1] http://webapp1.dlib.indiana.edu/vwwp/welcome.do

[2] https://perswww.kuleuven.be/~u0044428/clmetev.htm

[3] http://www.britannica.com

98K Britannica's articles in the three categories: introductory, intermediate and standard (Encyclopedia) and subject them to part-of-speech (POS) analysis. We have then determined significant differences between average numbers of POS tags across different categories (see Figure 1). We have also noticed higher number of verb phrases (VP) than prepositional phrases (PP) in easier article categories. Equation 1 shows the way to calculate POSR. High values point to easy texts.

$$\text{POSR} = \text{POS-Score} * \text{Chunk-Score}$$

$$\text{POS-Score} = (\#\text{NNS} + \#\text{JJS} + \#\text{NNPS} + \#\text{VB} + \#\text{VBD} + \#\text{VBP} + \#\text{VBZ}) / (\#\text{NN} + \#\text{VBG} + \#\text{VBN}) \tag{1}$$

$$\text{Chunk-Score} = \#\text{VP} / \#\text{PP}$$

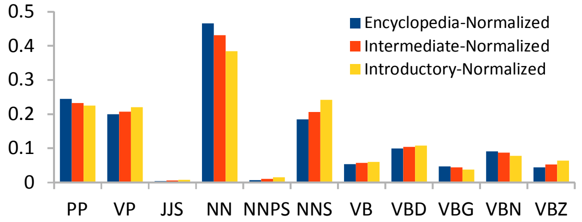

Figure 1 Average distribution of POS tags used in POSR measure across different categories of Britannica.

3. EXPERIMENTS AND CONCLUSIONS

Figures 2 and 3 demonstrate the outcomes of the tree readability measures in different decades for both the datasets. Fitted linear functions with R^2 values are overlaid on the figures. For the VWWPC dataset we can observe the steady increase in the ease of reading over time as measured by FRE (51% change) and by the tag-based POSR measure shown at the bottom of Figure 2 (57% change). The CRI index decreases along time, which again indicates improving ease of reading (-39% change). These results point to the apparent simplification of written language over time. The analysis of CLMETEV dataset (see Figure 3) confirms the results obtained for VWWPC; although the change seems to be less dramatic, especially for the POSR metric. This may be due to more diverse nature of this dataset.

Based on the results we conclude that recent texts are more readable to the current users in terms of the readability metric based on sentence and word lengths than the older texts on the basis of CLMETEV and VWWPC datasets. We have quantified the correlation of the two readability measures, FRE and CLI, and the text age. However, the readability metric using POS distribution gives a mixed picture. We have also observed varying levels of difficulty of texts written by different authors and published at the same decade. This justifies the need for multiple studies of readability using diverse corpora. We plan also to conduct detailed studies on the effect of language evolution on the comprehensibility of documents.

4. ACKNOWLEDGMENTS

This research was supported by MEXT Grant-in-Aid for Young Scientists B (#22700096) and JST research promotion program Sakigake: "Analyzing Collective Memory and Developing Methods for Knowledge Extraction from Historical Documents".

5. REFERENCES

[1] Danielson, W. A and Lasorsa, D. L. A New Readability Formula Based on the Stylistic Age of Novels. Journal of Reading, 33(3), 194-197, 1989

[2] Sherman, L.A., Analytics of Literature: *A Manual for the Objective Study of English Prose and Poetry*, Boston-Ginn, 1893.

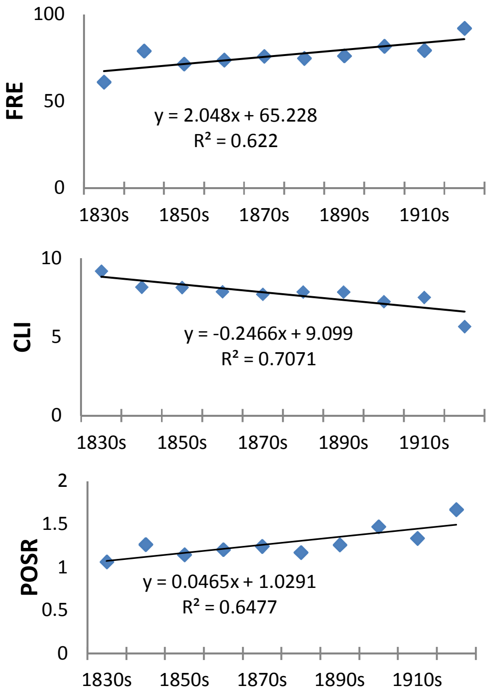

Figure 2 FRE, CLI and POSR readability measures of VWWPC dataset (σ = 3.39, σ = 1.05 and σ = 0.57, respectively).

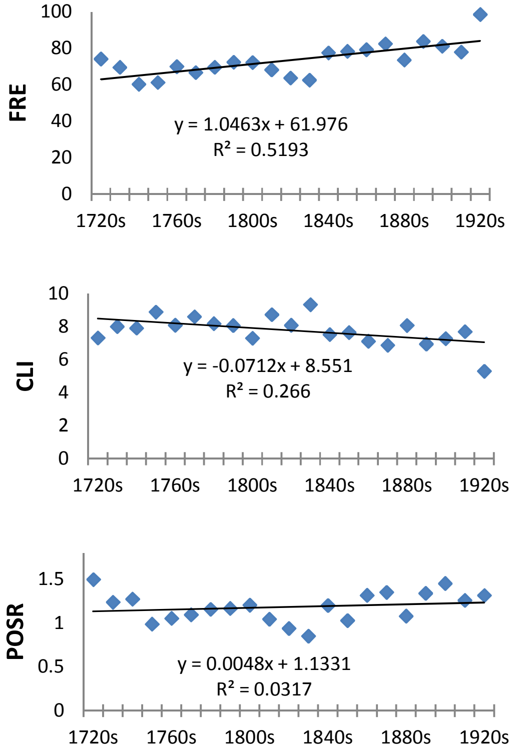

Figure 3 FRE, CLI and POSR readability measures of VWWPC dataset (σ = 3.15, σ = 1.07 and σ = 0.54, respectively).