Predicting Text Readability with Personal Pronouns

Boyang Sun, Ming Yue

Predicting Text Readability with Personal Pronouns

Boyang Sun[1] and Ming Yue[1]

[1]School of International Studies, Zhejiang University, Hangzhou, P.R. China
yueming@zju.edu.cn

Abstract. While the classic Readability Formula exploits word and sentence length, we aim to test whether Personal Pronouns (PPs) can be used to predict text readability with similar accuracy or not. Out of this motivation, we first calculated readability score of randomly selected texts of nine genres from the British National Corpus (BNC). Then we used Multiple Linear Regression (MLR) to determine the degree to which readability could be explained by any of the 38 individual or combinational subsets of various PPs in their orthographical forms (including ***I***, ***me***, ***we***, ***us***, ***you***, ***he***, ***him***, ***she***, ***her*** (the Objective Case), ***it***, ***they*** and ***them***). Results show that (1) subsets of plural PPs can be more predicative than those of singular ones; (2) subsets of Objective forms can make better predictions than those of Subjective ones; (3) both the subsets of first- and third-person PPs show stronger predictive power than those of second-person PPs; (4) adding the article ***the*** to the subsets could only improve the prediction slightly. Reevaluation with resampled texts from BNC verify the practicality of using PPs as an alternative approach to predict text readability.

Key Words: Readability, Personal Pronouns, Linear Regression

1. Introduction

The history of predicting textual readability quantitatively dates back to the 1940s when several linguists including Rudolf [1], George [2], Dale and Chall [3] introduced readability formulas into the field of research, thus unleashing a wave of researches and applications. Until 2017, Web of Science has published more than 11,000 researches on readability and its applications have moved from the field of education to fields of administration, commerce, computers, military, scientific research, etc. [4-6].

Traditional readability studies usually start with vocabulary and sentence complexity. For instance, the most widely recognized Flesch Reading Ease Formula uses word length (in terms of syllable) and sentence length (in terms of word count) as variables to calculate readability; the Dale-Chall Readability Formula exploits numbers of words that are not in the Dale-Chall 3000 Vocabulary and sentence length as criteria for predicting readability; the Gunning Fog Formula [7] and the SMOG Formula [8] employ number of polysyllabic words and sentence length as measures of readability. As computer technologies improve, many other factors are taken into account, such as type-token ratio, numbers of affixes, prepositional phrases and clauses, cohesive ties, other linguistics features [9], and even L2 learner's reading experience, etc. [10]. While these studies are valuable and significant, they usually involve multiple indirect indices that are subjectively defined or difficult to calculate in large-scale analysis. For example, it is hard to tell whether a word such as *factory* with two or more phonetic variants should be counted as 2 syllables (*/'fæktrɪ/*) or 3 syllables (*/'fæktəri/*). Besides, most of the classic formulae target for texts in English (and some other syllabic language), their applicability for non-syllabic languages such as Chinese remain untested.

In this research, we hope to test whether Personal Pronouns (hereinafter referred to as PPs) alone can have any predictive power for readability or not. There are several reasons for us to try them: (1) Given that PPs are always monosyllabic words used to replace full personal names or noun phrases, their usage in a text would affect its total word number, average sentence length as well as average word length; (2) PPs are often anaphorically used and can thus serve as cohesive ties to reduce redundancy and improve comprehension; (3) PPs were only tested collectively in [11] and [12] as part of linguistic features or cohesive ties, and consequently reached different conclusions on the role PPs play in readability prediction.

Since most languages have pronouns, we therefore propose that PPs could be promising candidate indicators of readability across languages and deserve further investigation. In this study, we will use a corpus-based approach to test the utility of individual PP forms in English texts of different genres. Specific research questions are as follows:

(1) Which person (first-, second-, or third- person, hereinafter referred to as 1P, 2P and 3P respectively) of PPs can predict text readability most accurately?

(2) Which number (Singular and Plural) of PPs can predict text readability more accurately?

(3) Which case (Subjective and Objective, with Possessive temporarily excluded)

of PPs can predict text readability more accurately?

Section 2 and Section 3 will introduce our research methods and data processing, Section 4 will report the data results from 5 aspects, Section 5 will reevaluate the results and Section 6 will summarize our major findings and limitations.

2. Materials and Methodology

This research uses corpus-based method and examines the predictability of various subsets of the PP forms (as shown in Table 1) on text readability in terms of Person, Number and Case.

It should be noted that the Possessive Case is not taken into consideration in this research. Nor will this paper look into the gender issue. So (***he***+***she***) and (***him***+***her***) will be considered as individual Subjective and Objective singular forms of 3P+HUMAN respectively; ***it*** be considered as the individual singular form of 3P-HUMAN with unclear Case; and ***you*** as the only 2P form with unclear Number and Case.

Consequently, there are 38 reasonable subsets of PP forms: 10 subsets with only individual PP forms, and 28 others with various Person/Number/Case combinations.

Table 1. Personal pronoun forms studied in this project

	1P		**2P**	**3P**		
	Singular	**Plural**	**Singular /Plural**	**Singular**		**Plural**
				+HUMAN	**-HUMAN**	
Subjective	*I*	*we*	*you*	*he + she*	*it*	*they*
Objective	*me*	*us*		*him + her*		*them*

2.1 Corpus data

British National Corpus (BNC) was chosen as our research object for the following reasons:

(1) All text materials in BNC were collected from native speakers as representative samples of Standard British English. So errors in pronoun use by non-native speakers have been excluded to a large extent; variations in geographical and social dialects should have been reasonably controlled or avoided as well.

(2) BNC contains approximately 100 million words, 90% of which are written materials collected from nine domains (also referred to as "genres" hereinafter) namely: (a) Arts; (b) Belief; (c) Commerce; (d) Imaginative; (e) Leisure; (f) World affairs; (g)

Natural science; (h) Social science; and (i) Applied science. Due to the different effects of genres on usages of PPs [13], proper sampling of this balanced general corpus allows for control over the genre variable that may affect readability.

Text materials used in this study (Corpus I) consist of 1,091,347 words in total, which are randomly selected from each of the nine domains. Corpus II consists of 972,490 words in total.

2.2 Readability Formula

In the present study, we choose the Flesch Reading Ease Score, which is recognized as the most widely used and the most tested and reliable formula [6], as approximants of real text readability to native readers. The specific formula is as follows:

$$Reading\ Ease\ Score\ =\ 206.835 - (1.015 * ASL) - (84.6 * ASW)$$

Where ASL = average sentence length (total word number divided by total sentence number), ASW = average word length (total syllable number divided by the total word number). The correlation coefficient between the Flesch readability formula and the Mc-Crabbs Reading Test was 0.7[1].

3. Data Processing

Data processing are divided into 4 steps:

(1) Use Perl program to count word and sentence length;

(2) Calculate the Flesch Reading Ease scores of sample texts of nine genres respectively;

(3) Use AntConc to count numbers of PP forms. Tokens of *US* as the abbreviation of the United States and tokens of the Possessive *her* are excluded during the retrieval. After that, the densities of the individual pronouns ($D_{(I)}$, $D_{(we)}$, etc.) based on the total word number of each text domain are calculated respectively;

(4) Use SPSS for multivariate regression analysis. Take the density of each subset of PPs as an independent variable, and the Flesch Reading Ease score as the dependent variable. Use Sig., correlation coefficient (R^2), as well as the adjusted correlation coefficient (adjusted R^2) values to determine which subset(s) of PPs may have better predictability. The criteria and process for determining moderate and strong fitting subsets are shown in Fig. 1.

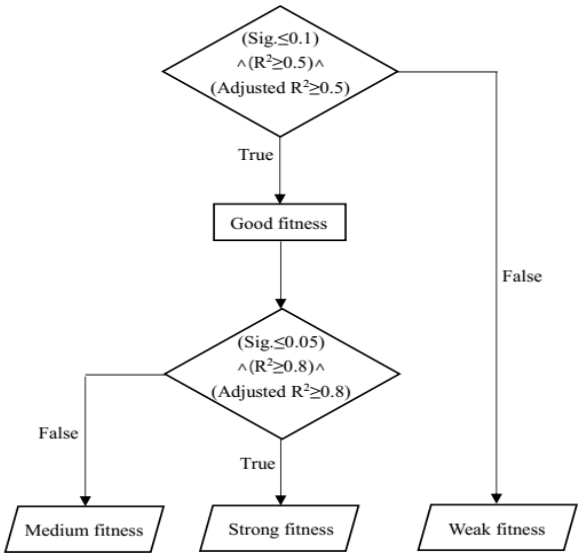

Fig. 1. Specific criteria for fitting degrees

4. Results and Discussion

4.1 Readability Results of Random Texts of Nine Genres

Table 2 shows that texts from Belief, Arts and Imagination domains are easiest to understand with highest readability scores among all texts from the nine domains; texts of Commerce, Natural Science, Applied Science and World Affairs are most difficult to read with lowest scores.

Table 2. Readability results for nine domains in BNC

Domain	Flesch Reading Ease Score	Difficulty Level
Belief	87.829	Easy
Arts	87.053	Easy
Imaginative	80.811	Easy
Leisure	67.623	Standard
Social science	51.449	Moderately difficult
Commerce	49.712	Difficult
Natural science	47.571	Difficult
Applied science	44.922	Difficult
World affairs	44.829	Difficult

4.2 Fitness Results

Individual Pronoun Forms and Readability. There are 10 subsets with individual PP forms as listed in Table 3. According to Fig. 1, it can be concluded that the subsets (***him*** + ***her***) and ***them*** present significant linear relations with readability and can explain almost 80% of variance ($R^2 \approx 0.8$), indicating strong predictive power. Additionally, ***us*** also shows a significant linear relation (Sig.=0.024) with accounting for about 50% of variance ($R^2=0.543$), showing that in contrast with individual Subjective PPs, individual Objective ones show fairly strong predictability on readability.

Table 3. Results for predictability of individual personal pronouns on readability

Pronoun forms	Case	Regression formulas	Sig.	R^2	Adjusted R^2
I	S	$R=1503.801*D_{(I)}+53.669$	0.091	0.354	0.261
we	S	$R=2328.992*D_{(we)}+53.662$	0.343	0.129	0.004
you	S+O	$R=2722.168*D_{(you)}+53.735$	0.102	0.336	0.241
he + she	S	$R=1208.909*D_{(he)}+728.189*D_{(she)}+52.435$	0.414	0.254	0.006
they	S	$R=11937.951*D_{(they)}+30.955$	0.101	0.338	0.243
me	O	$R=6757.564*D_{(me)}+54.975$	0.076	0.383	0.295
US	O	$R=14042.402*D_{(us)}+50.635$	**0.024**	**0.543**	**0.478**
him + her	O	$R=25621.555*D_{(him)}-10026.320*D_{(her)}+39.606$	**0.013**	**0.768**	**0.690**
them	O	$R=34550.512*D_{(them)}+13.615$	**0.000**	**0.863**	**0.843**
it	S+O	$R=4275.486*D_{(it)}+25.507$	0.064	0.408	0.324

Note: S stands for the Subjective Case; O for the Objective case. D() for word density in the text.

Person and Readability. The 38 individual and combinational subsets of PPs can be divided into seven groups according to Person (1P: 9 subsets; 2P: 1 subset; 3P: 12 subsets; 1P+2P: 1 subset; 1P+3P: 11 subsets; 2P+3P: 2 subsets; 1P+2P+3P: 2 subsets).

Results in Fig. 2 show that the 3P group has the best fitting degrees, with 5 subests (over 40%) of strong fitting and 2 (nearly 10%) of medium fitting subsets. The mixed (1P+3P) group performs similarly well, with 3 subsets (nearly 30%) of strong fitting and another 2 (nearly 10%) of good fitting subsets, way better than 1P and 2P subsets do. Therefore, it can be concluded that 3P subsets perform better than 1P and 2P subsets do in both individual and mixed subsets, which means that adding 1P and 2P subsets

into the 3P subsets will lowered their predictability.

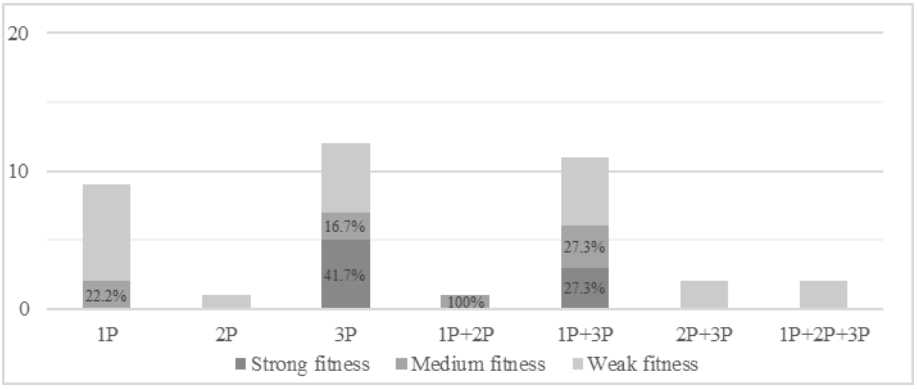

Fig. 2. Results for predictability of different Persons on readability in Corpus I

Number and Readability. The 38 individual and combinational subsets of PPs can be divided into three groups according to Number (singular PPs: 12 subsets, plural PPs: 9 subsets, singular + plural PPs: 17 subsets).

Fig. 3 shows that 50% of the singular-Number group offer good predication (with strong and/or medium fitness); and nearly 45% (11.1%+33.3%) of the plural-Number group show good prediction. The mixed-number group performs not as well.

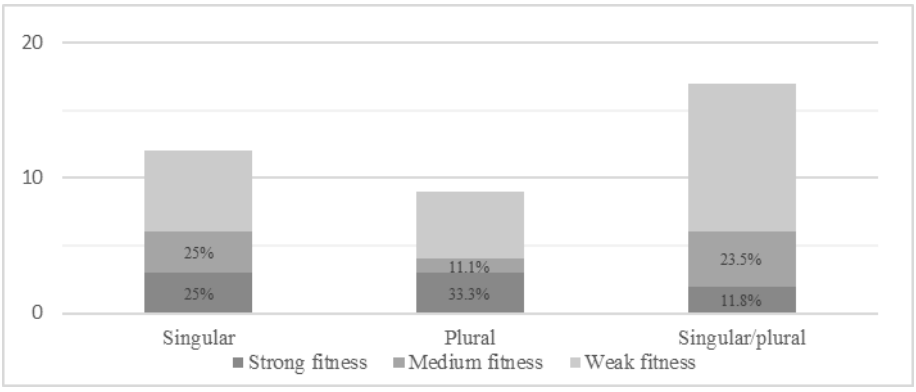

Fig. 3. Results for predictability of different Numbers on readability in Corpus I

Case and Readability. The 38 individual and combinational subsets of PPs can be divided into three groups according to Case (Subjective PPs: 9 subsets; Objective PPs: 9 subsets; Subjective + Objective PPs: 20 subsets).

Fig. 4 shows that Objective PP group has much stronger predictability than the Subjective group and the mixed-Case group, in both good and strong fitting area.

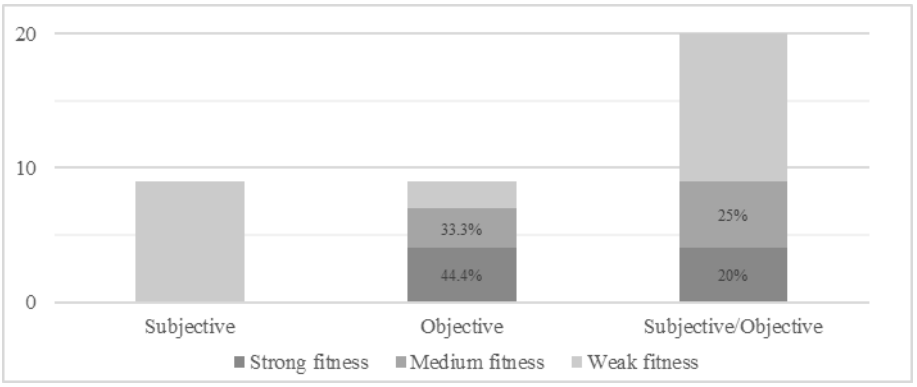

Fig. 4. Results for predictability of different Cases on readability in Corpus I

The* and Readability.** Since the definite article ***the in English has similar deictic/specifying function as pronouns do, we will test and see if this particular word and its combination with some of the PP subsets would have any predictive power on readability.

First, we use $D_{(the)}$ to predict text readability and gain a medium performance (Sig.=0.019, R^2=0.570, Adjusted R^2=0.509). Results in Fig. 5 show that subsets with ***the*** included perform slightly better than those without ***the*** in good and in strong fitting ranges. To test whether there is a significant difference while adding ***the*** in PPs, we use chi-square tests and draw the conclusion that the improvement is not significant (Chi-square value=0.213, df=2, p=0.899>0.05).

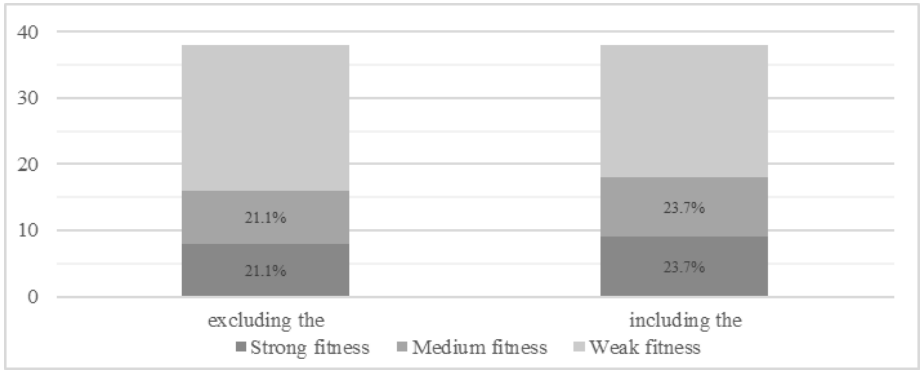

Fig. 5. Results for predictability of including and excluding ***the*** on readability

5. Reevaluation for Strong Fitting Subsets

All the subsets with a strong fitting degree are shown in Table 4. To explore whether subsets with strong predicting power can perform consistently, we repeated the procedures in Section 3 with re-sampled texts from BNC (Corpus II) and recalculated the pronoun and readability data in the new corpus. Test results from both Corpus I and II are shown in Table 4.

Table 4 shows that there are still two subsets with strong fitting degree in Corpus II, namely *"**he** + **him** + **she** + **her** + **it***" and ***"I** + **me** + **he** + **him** + **she** + **her** + **it"***. Although the other subsets have some changes in the fitting degree, they are almost in the moderate fitting range, indicating fair predictability.

Table 4. Personal pronoun subsets with strong fitness in Corpus I and II

Personal pronoun subsets	Corpus I			Corpus II		
	Sig.	R^2	Adjusted R^2	Sig.	R^2	Adjusted R^2
them	0.000	0.863	0.843	0.070	0.670	0.623
us+them	0.002	0.871	0.827	0.036	0.670	0.560
him+her+them	0.001	0.959	0.934	0.110	0.872	0.796
me+us+him+her+them	0.024	0.962	0.899	0.128	0.879	0.677
he+him+she+her+it	**0.022**	**0.964**	**0.905**	**0.005**	**0.986**	**0.963**
I+me+he+him+she+her+it	**0.021**	**0.999**	**0.999**	**0.036**	**0.999**	**0.998**
he+him+she+her	0.004	0.964	0.928	0.036	0.887	0.774
they+them	0.002	0.869	0.825	0.035	0.672	0.563

6. Conclusion

A corpus-based approach is used in research to explore the readability predictability of 77 subsets with various personal pronoun forms and the definite article ***the***. The results show that: (1) *them* has the best predictive power among individual pronoun forms; (2) 3P and 1P make better predictions than 2P; (3) plural PPs outperforms singular ones only in strong fitting range; (4) Objective PPs can predict more accurately than Subjective ones; (5) definite article ***the*** may only improve subsets' predictability slightly; (6) Retesting results are consistent for those PP subsets with good predictability. Therefore, we believe that using specific subsets of PPs to predict text readability appears practical.

However, large-scale tests are needed before any solid conclusion can be drawn concerning the applicability of PPs for readability prediction. Detailed investigation

into the predictability of Possessive PPs, and ***it*** in Subjective and Objective Cases may be needed as well. Besides, it needs to be verified on whether texts in other geographical varieties such as American English are similar to their British matches.

References

1. Flesch R.: A new readability yardstick. *Journal of Applied Psychology* 32, 221-233 (1948).
2. Klare G.R.: Measures of the readability of written communication: An evaluation. *Journal of Educational Psychology* 43(7). 385-399 (1952)
3. Dale E., Chall J.S.: A formula for predicting readability: instructions. *Educational Research Bulletin* 2(27), 37-54 (1948).
4. Meppelink C.S., van Weert J.C.M., Brosius A., Smit E.G.: Dutch health websites and their ability to inform people with low health literacy. *Patient Education and Counseling* 11(100), 2012-2019 (2017).
5. Botas S., Veiga R., Velosa A.: Bond strength in mortar/ceramic tile interface-testing procedure and adequacy evaluation. *Materials and structures* 2115(50), (2017).
6. Dubay W.H.: Smart language: readers, readability and the grading of text. Costa Mesa, California (2007)
7. Gunning R.: The technique of clear writing. McGraw-Hill (1952).
8. Laughlin G.H.M.: SMOG grading-a new readability formula. *Journal of Reading* 12(8), 639-646 (1969).
9. Pitler E., Nenkova A.: Revisiting readability: a unified framework for predicting text quality. EMNLP (2008).
10. Kotani, K., Yoshimi, T.: Measuring readability for learners of English as a foreign language by linguistic and learner features. In Hasida, K., Purwarianti, A. (eds.), Computational Linguistics: 14th International Conference of the Pacific Association for Computational Linguistics, PACLING 2015, Bali, Indonesia, May 19-21, 2015, Revised Selected Papers, pp211-222. Springer Singapore, Singapore (2016).
11. Brinton J.E., Danielson W.A.: A factor analysis of language elements affecting readability. *Journalism Quarterly* 35(4), 420-426 (1958).
12. Todirascu A., Ois T.F., Bernhard D.: Are cohesive features relevant for text readability evaluation? 987-997 (2017).
13. Biber D.: Spoken and written textual dimensions in English: resolving the contradictory findings. *Language* 2(62), 384-414 (1986).