Analyzing the Readability Statistics of USPTO Patent Grants

Eren Orbey[1]

[1]Department of Computer Science, Yale College, New Haven, CT

LILY Lab

Introduction

Technological innovation in the last decade has contributed to an increasing digitization of human interaction and communication in the form of text. Researchers, in turn, have introduced text as an input to data analysis across economic sectors. Governments, businesses, and educators across the world have longed relied on so-called readability scores as objective proxies for the complexity of English-language textual documents. These formulas have provided minimum thresholds of clarity and comprehensibility for financial disclosure forms, standardized test questions, technical manuals, and medical inserts. Higher readability scores correlate in some sectors with increased readership. (More readable articles in academic journals tend to be cited more frequently and to win more awards.) This project seeks to apply common readability metrics to the text of archives from the United States Patent and Trademark Office (USPTO) to determine whether and how the clarity and concision of patents vary across industry and time.

Materials and Methods

I focus on four of the most widely used and tested readability scores—the Flesch Reading Ease, Flesch-Kincaid, Simple Measure of Gobbledygook (SMOG), and Dale-Chall metrics—to assess differences in clarity and concision in the abstracts of USPTO patent grants from 2006 to 2015. In order to standardize this analysis, I adhere to the hierarchical International Patent Classification (IPC) system, grouping grants according to their industry designations, which may be liable to increase or decrease the overall readability of an entire classification of patents. I also consider the time elapsed between the filing and the publication of each patent to determine whether and how the clarity and concision of grants affect their passage toward approval by the USPTO bureaucracy. The purpose of this approach is to investigate the efficacy of classifying patents according to the comprehensibility of their grant text. Is there an incentive for inventors to file clearer, less complex patent grants—even though clarity might decrease the barrier to entry for future competitors?

Year	Mean Readability Scores				Mean Time to Publication (Days)
	Flesch	Flesch-Kincaid	SMOG	Dale-Chall	
2006	32.05	18.50	16.28	11.69	1160
2007	31.26	18.72	16.39	11.76	1196
2008	30.31	18.95	16.48	11.82	1255
2009	29.62	19.24	16.59	11.87	1330
2010	29.39	19.28	16.64	11.90	1383
2011	28.22	19.67	16.79	11.96	1351
2012	27.72	19.81	16.86	11.99	1315
2013	27.08	20.00	16.95	12.03	1256
2014	26.84	20.08	16.98	12.05	1244
2015	26.68	20.17	17.00	12.06	1205

Table 1. Mean Readability (Flesch, Flesch Kincaid, Smog, and Dale-Hall Scores) by Year (2006 to 2015). Recall that more readable texts produce higher Flesch scores and lower Flesch- Kincaid, SMOG, and Dale-Chall grade levels—that is, the small *decrease* in mean Flesch scores and smaller *increase* in the others demonstrates a slight decline in patent readability. The texts decline in readability over the decade.

IPC Class (A–H)	Mean Readability Scores				Mean Time to Publication (Days)
	Flesch	Flesch Kincaid	SMOG	Dale-Chall	
A	32.83	17.72	16.13	11.79	1335
B	31.94	19.69	16.10	11.24	1223
C	24.62	20.04	17.73	13.05	1366
D	31.34	19.86	16.20	11.40	1278
E	39.08	17.30	15.11	10.79	1151
F	34.43	18.88	15.65	11.15	1200
G	26.11	19.99	17.19	12.16	1295
H	26.94	20.02	16.91	12.02	1235

Table 2. The readability scores rank Classes C (Chemistry; Metallurgy), G (Physics), and H (Electricity) as consistently more difficult texts. The correlation between readability and time to publication is not definitive across IPC classes— consider that Class A (Human Necessities) shows a high time to publication despite being relatively readable on average. It is notable, though, that Class C, statistically the least readable, shows the high mean time to publication.

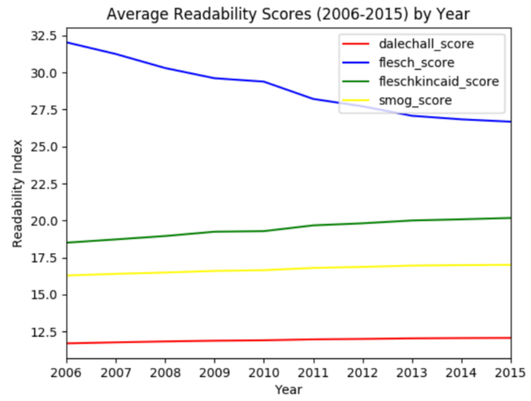

Figure 1. Mean Readability (Flesch, Flesch Kincaid, Smog, and Dale-Hall Scores) by Year (2006 to 2015).

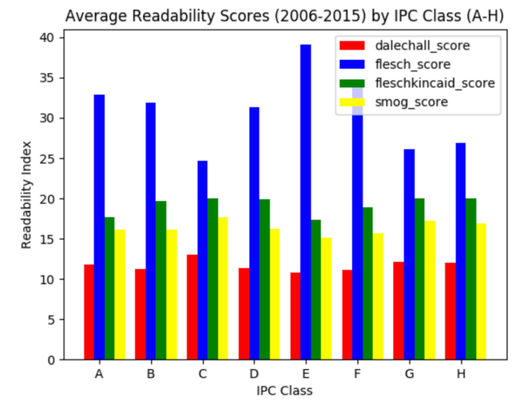

Figure 2. Mean Readability (Flesch, Flesch Kincaid, Smog, and Dale-Hall Scores) by IPC Class (A-H).

Results

The data suggest a decline in patent readability between 2006 and 2015, as demonstrated by increasing mean Flesch indices and decreasing mean Flesch-Kincaid, SMOG, and Dale-Chall grade levels (Table 1, Figure 1). Patents assigned to IPC Classes C (Chemistry; Metallurgy), G (Physics), and H (Electricity) rank consistently as the most difficult, least readable texts according to all four metrics and correspond to longer delays between application and publication (Table 2, Figure 2). (The least readable patent class, in particular, corresponds to the highest mean time to publication.) Within each IPC Class (A-H), the data do not conclusively indicate a direct or inverse variation between calculated readability scores and time required for publication. Further experimentation ought to be performed within subsets of specific classes to determine the predictive capacity of readability scores.

Conclusion

I introduce the question of whether readability predicts the speed at which filed patents are approved and published—a crucial metric for competitive inventors racing to lay claim to intellectual property. Though the data do not determine a conclusive relationship between these variables within each IPC Class, the differences in readability between classes, and the corresponding differences in time to publication, provide the motivation to conduct further experimentation by applying the existing script to assess a broader span of patent data and/or by modifying the existing script to consider patent classifications more granular than the top-level IPC categories.

Acknowledgements

I would like to thank Dragomir Radev for his time, insight, and generosity.