ProtAnt: A freeware tool for automated prototypical text detection

Laurence Anthony
Waseda University
anthony@waseda.jp

Paul Baker
Lancaster University
j.p.baker@lancaster.ac.uk

1 Introduction

Prototypicality can be defined as "having the typical qualities of a particular group or kind of person or thing" (Merriam-Webster 2014). In quantitative and qualitative corpus-based studies, researchers are often interested in identifying prototypical texts so that they can conduct close readings and begin to examine the 'why' behind the numbers revealed through top-down, broad-sweep quantitative analyses. Researchers in other areas may also need to identify prototypical texts in order to, for example, classify texts according to genre, locate typical student essays at a particular level for instructional purposes, flag texts (e.g. extremist writing) for further analysis, or remove outlier texts from a corpus before conducting a quantitative study.

In this paper, we present a novel approach to prototypical text detection that is fast, completely automated, and statistically rigorous. Our method does not require manual assignment of texts to pre-conceived classes as is the case with many natural language processing methods, and it is able to rank texts by their prototypicality in a way that is meaningful and easy to interpret. We have encapsulated our approach in a free software tool, *ProtAnt*, that runs on Windows, Macintosh OS X, and Linux operating systems, and is designed to be easy-to-use and intuitive even for novice users of computers.

2 The *ProtAnt* approach

The starting point for our prototypical text detection approach is to identify key words in a corpus. Key words are 'words' that appear statistically significantly more frequently in the target corpus than in a suitable reference corpus. Depending on the design of the target corpus and choice of reference corpus, these key 'words' may be lexical items, part-of-speech tags, discourse moves, or a multitude of other linguistic features that can be coded or annotated. For this study, we focus on lexical (word) prototypicality. Our *ProtAnt* tool detects these key words using a standard log-likelihood statistically measure of keyness (Dunning 1993), but other measures can be easily incorporated.

The second stage in our approach is to rank the key words so that the most salient key words can be selected for use in further analysis, and the least salient key words removed. Almost all previous corpus-based studies utilizing key words have ranked the words based on the raw 'keyness' value as given by the statistical measure (e.g. log-likelihood). This is equivalent to ranking the words by their p-value. A more informed way to rank key words is by considering the (normalized) size of difference in frequency between the target and reference corpus, i.e., the key word's effect size. There are many ways this can be calculated, including relative frequency (Demarau 1993) or a log of relative frequency (e.g. Hardie 2014). In *ProtAnt*, the user is given a choice of ranking key words by either p-value or effect size measures.

The final stage in the *ProtAnt* approach is to count the number of key words in each corpus file, normalize the counts by the length of the texts, and then rank the corpus texts by the number of key words they contain. Texts containing high numbers of key words are those that contain more words that characterize the corpus as a whole and thus can be considered to be prototypical of the corpus as a whole.

Figure 1 shows a screenshot of the *ProtAnt* tool after completing an analysis of a small corpus of 20 newspaper articles using the BE06 Corpus (Baker 2009) as a reference corpus. In the screenshot, the top right table shows that file 7 is the most prototypical. The middle table shows the key words contained in each file, with file 7 shown to include the words "islam," "blair," "muslim," "brotherhood" and other topic related words. The bottom table shows a complete list of the key words, here created by log-likelihood and ranked by p-values.

3 Validation experiments

Five experiments were conducted to establish the validity of the *ProtAnt* approach to prototypical text identification. The first experiment was designed to see if *ProtAnt* was able to correctly identify prototypical texts in a small corpus of newspaper articles. For this experiment, the corpus was artificially designed to contain 10 texts on the topic of Islam (deemed to be the main theme of the corpus), 5 texts related to the general topic of football (serving as a distractor theme), and 5 texts with no overlapping topics of focus, with the BE06 corpus serving as a reference corpus. A successful *ProtAnt* analysis should be able to rank the 10 texts on Islam higher than the other texts.

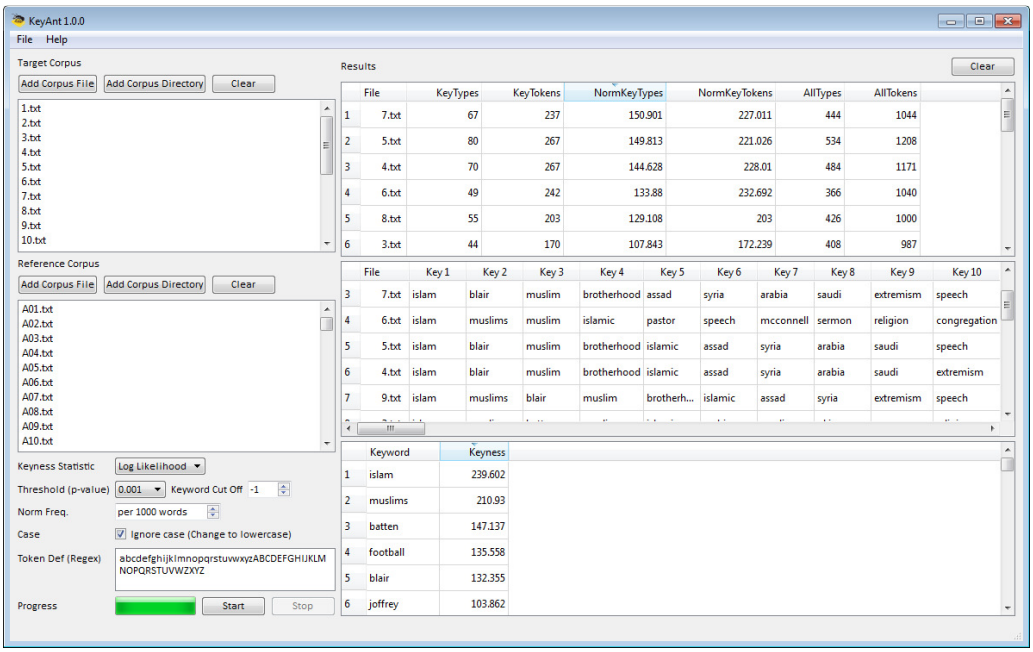

Figure 1: Screenshot of the *ProtAnt* prototypical text detection tool

Table 1 shows the results of the *ProtAnt* analysis for a log-likelihood (LL) threshold value of 0.001 with the texts rank ordered by normalized key type and normalized key token values. Clearly, the *ProtAnt* analysis was able to reliable rank almost all Islam files as the most prototypical of the corpus as a whole, regardless of whether key types or key tokens are used. The rankings were also shown to be stable regardless of the log-likelihood threshold value. Interestingly, one of the Islam texts was unexpectedly ranked lower in the lists. A close reading of this text, however, revealed several unusual features that were not immediately apparent to the investigators; it is a story about a school which told parents that children had to attend a workshop on Islam or be called racist. Thus, this ranking serves as further evidence of the usefulness of the *ProtAnt* tool.

	LL threshold (0.001)	
Rank	Key Types	Key Tokens
1	Islam	Islam
2	Islam	Islam
3	Islam	Islam
4	Islam	Islam
5	Islam	Islam
6	Islam	Islam
7	Islam	Football
8	Islam	Obituary
9	Islam	Islam
10	Football	Islam
11	Obituary	Islam
12	Islam	Football
13	Review	Science
14	Football	Review
15	Science	Islam
16	Tennis	Tennis
17	Football	Football
18	Football	Art
19	Football	Football
20	Art	Football

Table I: *ProtAnt* analysis of newspaper articles

The second experiment was designed to see if *ProtAnt* was able to correctly identify prototypical texts in a small corpus of longer novels. Following a similar design to that used in experiment 1, 10 versions of the novel *Dracula* were compared against five versions of the novel *Frankenstein*, and 5 other randomly selected novels. Again, results revealed that the *ProtAnt* analysis could rank almost

all *Dracula* texts above the other novels in the corpus, with the results remaining stable regardless of key type or key token ordering, or choice of log-likelihood threshold value (results not shown).

Experiments 3 and 4 were designed to see if *ProtAnt* could identify prototypical texts in a larger, traditional corpus. For experiment 3, we performed a *ProtAnt* analysis of texts in the AmE06 Corpus (Potts and Baker 2012) using the BE06 corpus as a reference corpus in order to find prototypical texts that are 'American' in nature. For experiment 4, we performed a *ProtAnt* analysis of texts in the AmE06 Corpus, but this time used the Brown Corpus (Francis & Kucera 1963) as a reference corpus in order to identify prototypical texts expressing the concept of 'the year 2006'. Again, convincing results from the *ProtAnt* analysis were obtained in both experiments, with the highest ranked texts clearly expressing the target themes. For example, in experiment 4, the highest ranked text was a fairly dry government text about tax. It is written with a direct address to the reader and makes frequent use of the second person pronoun key words *you* and *your* (a feature of personalizing language that has become more popular since 1961).

Experiment 5 was designed to see if the *ProtAnt* analysis was able to find outliers in a corpus. For this experiment, we again used AmE06 (with BE06 as the reference), but this time selected all the files from one register and artificially added an additional file randomly selected from a different register. A successful analysis should rank the artificially added file as the lowest in the list. When the experiment was repeated for all registers in AmE06, results showed that the outlier file could be correctly identified as being at the bottom or very close to the bottom of the list (within 2) in 10 out of the 15 cases.

4 Conclusion

In this paper, we have shown that a prototypical text detection approach based on ranking texts according to the number of key words they contain can be successfully applied in a variety of test-case situations. We have also developed a software tool that allows researchers to apply the approach as part of their own analysis through an easy-to-use and intuitive interface. Our software tool, *ProtAnt*, is freely available at the following URL: http://www.laurenceanthony.net/software.html. We hope this tool will introduce traditional qualitative researchers to the advantages of corpus-based approaches, and also remind quantitative corpus-based researchers of the importance of close readings of corpus texts.

References

Baker, P. 2009. “The BE06 Corpus of British English and recent language change”. *International Journal of Corpus Linguistics* 14(3): 312-337.

Damerau, F. J. 1993. “Generating and evaluating domain-oriented multi-word terms from texts”. *Information Processing and Management* 29: 433-447.

Dunning, T. 1993. “Accurate Methods for the Statistics of Surprise and Coincidence”. *Computational Linguistics* 19(1): 61-74.

Francis W. N. and Kucera H. 1964. Brown Corpus. Available online at https://archive.org/details/BrownCorpus

Merriam-Webster. 2014. Available online at http://www.merriam-webster.com/dictionary/prototypical

Potts, A. and Baker. P. 2012. “Does semantic tagging identify cultural change in British and American English?” *International Journal of Corpus Linguistics* 17(3): 295-324.