Cluttered Writing: Adjectives and Adverbs in Academia

Adam Okulicz-Kozaryn*
Rutgers - Camden

Draft: Sunday 9th December, 2012

Abstract

Scientific writing is about communicating ideas. Today, simplicity is more important than ever. Scientist are overwhelmed with new information. The overall growth rate for scientific publication over the last few decades has been at least 4.7% per year, which means doubling publication volume every 15 years. I measure simplicity/readability with proportion of adjectives and adverbs in a paper, and find natural science to be the most readable and social science the least readable.

"When you catch an adjective, kill it."
Mark Twain

"The road to hell is paved with adverbs."
Stephen King

Scientific writing is about communicating ideas. Clutter doesn't help–texts should be as simple as possible. Today, simplicity is more important than ever. Scientist are overwhelmed with new information. The overall growth rate for scientific publication over the last few decades has been at least 4.7% per year, which means doubling publication volume every 15 years (Larsen and von Ins 2010). How do we keep up with the literature? We can use computers to extract meaning from texts (Hopkins and King 2007). Better yet, I propose here, we should be writing research in machine readable format, say, using Extensible Markup Language (XML). I think, it is the only way for scientists to cope with the volume of research in the future. But the first step is to start writing as simply as possible to minimize the volume and maximize the meaning. Readability of scientific writing matters not only for scientists. Readable scientific writing could reach wider audience and have a bigger impact outside of academia.

So how do we produce readable and clean scientific writing? One of the good elements of style is to avoid adverbs and adjectives (Zinsser 2006). Adjectives and adverbs sprinkle paper with unnecessary clutter. This clutter does not convey information but distracts and has no point especially in academic writing, say, as opposed to literary prose or poetry. William Zinnser, one of the writing experts, advises (Zinsser 2006):

> Most adverbs are unnecessary. You will clutter your sentence and annoy the reader if choose a verb that has a specific meaning and then add an adverb that carries the same meaning[...] Most adjectives are also unnecessary. Like adverbs they are sprinkled into sentences by writers who don't stop to think that the concept is already in the noun. This kind of prose is littered with precipitous cliffs and lacy spiderwebs[...]

Why measuring readability by counting adjectives and adverbs? There are many readability measures, for instance: Gunning Fog Index, Automated Readability Index, Coleman-Liau Index, Flesch-Kincaid Reading Ease, Flesch-Kincaid Grade Level, SMOG Index, FORCAST Readability Formula. They are based on counts of words, difficult words (many syllables), and sentences. And the calculated measure is usually a grade level required to understand the text. I did not use these measures for two reasons. First, to calculate these indices I would need full texts of published research, and it appears that as of 2012 I cannot bulk download enough full texts to have a representative sample of a discipline. Second, counting syllables is not a trivial task, and it appears that there are many ways to do it, and the software is not very mature.

*EMAIL: adam.okulicz.kozaryn@gmail.com All mistakes are mine.

At the same time, adjectives and adverbs counts are a relatively useful measure. They can be calculated using mature NLTK module for Python (Bird 2006). NLTK (Natural Language Toolkit) is a module for Python programming language that can be used for analysis of human language, for instance, to calculate proportion of adjectives and adverbs in text. Both NLTK and Python are free and run on Linux, Mac, and Windows. They can be downloaded from `http://nltk.org/` and `www.python.org/`. You will also find extensive documentation and tutorials at the above addresses. NLTK comes with a number of dictionaries that can be used to identify parts of speech, say adjectives and adverbs.

I use data from JSTOR Data For Research (`http://dfr.jstor.org/`). The sample is about 1,000 articles randomly selected from all articles published in each of seven academic fields between 2000 and 2010. I made the following selection from JSTOR:

1. Content Type: Journal (to analyze research, not the other option: Pamphlets)
2. Page Count: [5 TO 100] (to avoid short letters, notes, and overly long essays; fewer than 5 pages may not offer enough to evaluate text, and longer than 100 may have a totally different style than the typical one for a given field)
3. Article Type: Research article (other types such as book reviews may contain lengthy quotes, etc)
4. Language: English
5. Year of Publication: [2000 TO 2010] (only recent research; did not select 2011, 2012, since for some fields JSTOR does not offer most recent publications–the number of available articles in most recent years dramatically drops, based on a JSTOR graph available at the selection)

I identify parts of speech using Penn Tree Bank in Python NLTK module (Bird 2006). I calculate the proportion of adjectives and adverbs for each academic discipline, and divide it by the smallest proportion, so that results show proportion increase over the discipline with the smallest proportion of the adjective-adverb clutter. Figure 1 shows that natural science uses the fewest adjectives and adverbs, while social science uses the most (about 15% more than natural science).

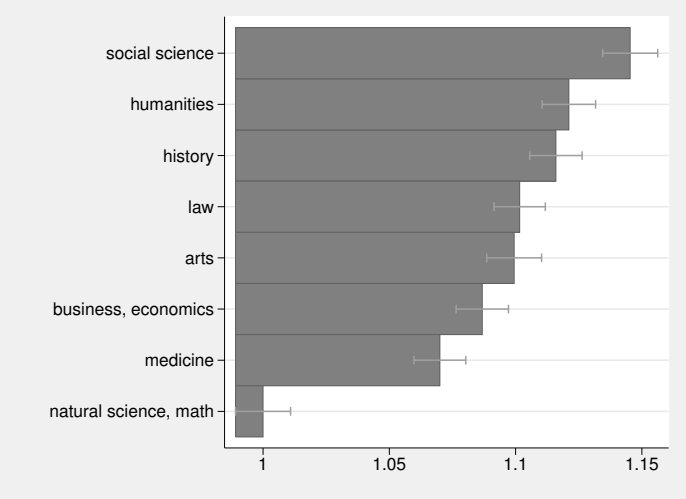

Figure 1: Proportion of adjectives and adverbs in published research by academic discipline group relative to the field with the smallest proportion. 95% confidence intervals shown.

Is there a reason that a social scientist cannot write as clearly as a natural scientist? Again, adjectives and adverbs are often meaningless and sometimes misleading. And there is a software to check for the proportions of parts of speech: Python's NLTK (Bird 2006). Following Mark Twain, the scientists should kill much of the adjectives and adverbs to make the academic prose readable and spare us from the unnecessary increase in the volume of research output.

References

Bird, S. (2006): "NLTK: the natural language toolkit," in *Proceedings of the COLING/ACL on Interactive presentation sessions*, Association for Computational Linguistics, 69–72.

Hopkins, D. and G. King (2007): "Extracting systematic social science meaning from text," *Manuscript available at http://gking. harvard. edu/files/words. pdf*.

Larsen, P. and M. von Ins (2010): "The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index," *Scientometrics*, 84, 575–603.

Zinsser, W. (2006): *On writing well: The classic guide to writing nonfiction*, Harper Paperbacks.