Cluttered Writing: Adjectives and Adverbs in Academia

Adam Okulicz-Kozaryn*
Rutgers - Camden

Draft: Monday 1st October, 2012

"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 (1). How do we keep up with the literature? We can use computers to extract meaning from texts for us (2). 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.

So how do we produce readable and clean scientific writing? One of the good elements of style is to avoid adverbs and adjectives (3). 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 (3):

> 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[...]

Research should be brief and to the point. And 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 does scientific writing score on the adjective-adverb clutter by discipline ?

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 identify parts of speech using Penn Tree Bank in Python NLTK module (4). I calculate the proportion of adjectives and adverbs for each academic discipline, and divide it by the smallest, 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.

*EMAIL: adam.okulicz.kozaryn@gmail.com

I am indebted to and. All mistakes are mine.

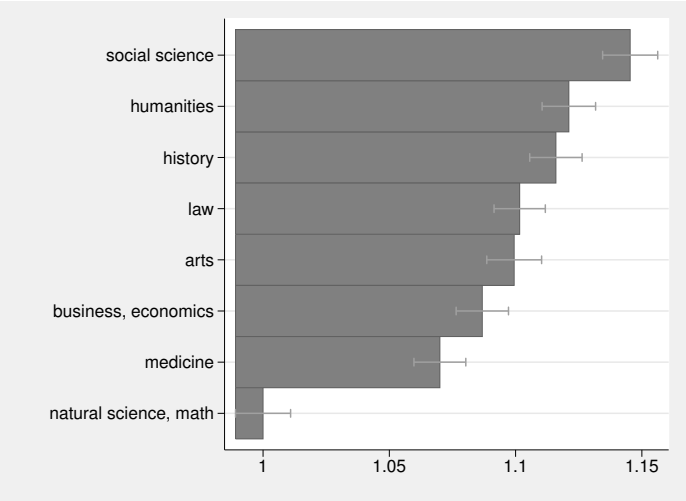

Figure 1: Proportion of adjectives and adverbs in published research by academic discipline group relative to the field with the smallest proportion. 95% confidence intervals shown.

Is there a reason that a social scientist writes less clearly than 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 (4). Following Mark Twain, the scientist should kill much of the adjectives and adverbs to make her academic prose readable and spare us from the unnecessary increase in the volume of research output.

References

[1] Larsen P.O., Ins M.. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index *Scientometrics*. 2010;84:575–603. readability_science.

[2] Hopkins D., King G.. Extracting systematic social science meaning from text *Manuscript available at http://gking. harvard. edu/files/words. pdf*. 2007. readability_science.

[3] Zinsser W.. *On writing well: The classic guide to writing nonfiction*. Harper Paperbacks 2006. readability_science.

[4] Bird S.. NLTK: the natural language toolkit in *Proceedings of the COLING/ACL on Interactive presentation sessions*:69–72Association for Computational Linguistics 2006. readability_science.

Technical appendix

Why counting adjectives and adverbs?

There are many readability measures, for instance: Gunning Fog Index, Automated Readibility 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 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.

At the same time, adjectives and adverbs counts are a relatively useful measure. They can be calculated using mature NLTK module for Python (4). And JSTOR provides word counts that can be used for this purpose.

The reason for reporting of results relative to the field with the smallest proportion of adjectives and adverbs is that the parts of speech classification is not 100% accurate by definition and, hence, the relative results are more meaningful.

JSTOR selection

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)

tagging

Tagging is not a straightforward task. There are many ways to do it. And by no means I am an expert at it. For robustness I used a default NLTK tagger, and then checked with a trained (on a full Brown corpus) unigram tagger. Here are results (proportions of adjectives-adverbs to other parts of the speech) for the default tagger and the unigram tagger trained on Brown corpus.

jou	default (pos)	unigram(Brown)
DisciplineGroup_Arts_00_10	12.6	9.6
DisciplineGroup_Business_and_Eco	12.4	10.2
DisciplineGroup_History_00_10	12.8	10.0
DisciplineGroup_Humanities_00_10	12.8	10.1
DisciplineGroup_Law_00_10	12.6	10.4
DisciplineGroup_Medicine_and_All	12.2	9.4
DisciplineGroup_Science_and_Math	11.4	8.7
DisciplineGroup_Social_Sciences_	13.1	10.5
Total	12.5	9.8

Results in both columns are substantively similar. The second column show smaller absolute proportions as expected: it tags only the words it saw in the training corpus.

code

Python and Stata code are in the separate documents.