Reporting and Editorializing: Linguistic Differences

Joseph J. Devney

The job of a newspaper is to report the news, and to try to report it objectively, without bias. But the entire paper does not aim for the ideal of objectivity. All major American newspapers include a small section called the "op-ed" section (for "opinion and editorial"). This is where personal opinions, as opposed to objective facts, appear. The opinions come from the paper's staff members (editorials), from outside commentators (opinion pieces), and from the paper's readers (letters to the editor). These items are necessarily different from the reportage in the rest of the paper. But how are they different? Are there linguistic differences that can be identified or measured? My linguistic examination of newspaper content found differences in the lexicon and writing style between the two kinds of newspaper content.

The corpora used for this study were subsets of the Brown Corpus. The Brown Corpus is subdivided into several genres. Two of these genres are *Press: Reportage* and *Press: Editorial*. Each of these is subdivided into smaller groupings. *Reportage* includes Political, Sports, Society, Spot News, Financial, and Cultural reporting. While each individual subgrouping might have its own vocabulary or preferred writing style, any linguistic analysis of the entire contents of the *Reportage* genre should give results that reflect the non-editorial voice of the paper as a whole. *Editorial* has three groups: institutional, personal, and letters to the editor. These are analogous to the editorials, opinion pieces, and letters. My goal for this project is to compare the two types of writing, to see what sort of differences there might be between the two.

At the level of individual words, the two categories were compared to determine if they differed in the vocabulary they use, and in how often particular word types appear. This is partly because the range of topics that appear on the op-ed pages will likely be smaller than for the rest of the paper. For example, few editorials concern sports, but most newspapers have an entire daily section devoted to sports. Politics and government are likely to dominate on the editorial pages.

Proper nouns will likely indicate the topics discussed in editorials versus reporting. Words relating to federal government affairs or international events are more likely to appear in editorials.

Adjectives with especially positive, negative, or emotional connotations might appear more often on the editorial page.

At the level of the clause, it is likely that the editorials and opinion pieces will use more relative clauses. These argumentative essays are written in a different register than are the news articles, more resembling an academic writing style than a bare-bones descriptive style.

For this reason, sentence length might also differ between the two categories. If there is a difference in average sentence length, sentences in straight reportage are likely to be shorter. Paragraphs in newspaper articles tend to be short, often only a single sentence. This may also be true for the sentences themselves. Writers of editorials and op-ed pieces, on the other hand, may need to spend some time and effort developing an idea and supporting it, which may make for more complex, and therefore longer, sentences.

One potential difficulty with using the Brown corpus is that it is not current American English. All the material I will be examining was printed in 1961, almost 50 years ago. Half a century may be long enough for changes in language and writing style to become apparent.

Certainly some of the editorial and news topics will be different. In 1961, we were at the height of the Cold War with the Soviet Union and China. For this reason, terms like *Atom bomb* and *Communism* will likely appear more often than they would in a current newspaper. And obviously a 21-century newspaper will include many references to technologies (and even countries!) that did not exist in 1961, including cell phones, the Internet, catalytic converters, and microchips. Some words, like fax and Walkman, have appeared and almost disappeared since 1961, along with the technologies they describe.

Methodology

Technical problems prevented me from doing the more thorough research I had planned at first. The corpora were not formatted for Tgrep2, and were not available for use with CQP. Therefore, I did what data extraction I could using UNIX commands.

In order to identify the subjects discussed in the op-ed section versus the ones reported on in the rest of the paper, the first step was to identify proper nouns. These would include the people who were quoted, names of places discussed, and organizations such as companies, government agencies and political parties. The proper nouns for each corpus were ranked by frequency (using lemmas), most frequent to least frequent.

Making sense of the results, however, is an exercise that goes beyond simply gathering statistics from the corpus, because the nouns are not broken down by meaning. Further analysis on the sense of the results must be performed by humans, to understand exactly what is meant by “Russia” or “Republican.”

Adjectives and adverbs were extracted separately, but analyzed in the same way. For each part of speech, the lemmas of the words that appeared in each corpus were ranked in order of frequency of appearance. Taking the first 50, those that had connotations of judgment were identified and counted. For example, simple reportage might say that a company spokesman “said” something. If, instead, it is reported that the spokesman “claimed” the same thing, this phrasing casts doubt on the truthfulness of the speaker.

Because the data available was not accessible to Tgrep2, the next step in the data gathering was not possible. This would have been to calculate how often relative clauses and other indications of complex sentences were used. The expectation is that the editorial page writers would use more complex sentences. As a surrogate for this more sophisticated measurement, I calculated the average sentence length in the two corpora.

Limitations of the Data

The portions of the Brown Corpus used for this research was not formatted for use with Tgrep2. It also was not available for use with CQP. All data extraction was done with UNIX command-line tools.

The POS tagging in the corpus caused some difficulty. A proper noun such as “General Assembly” (of the United Nations) was coded in this way:

```
[ The/DT General/NNP Assembly/NNP ]
```

For purposes of analyzing the topics discussed in the material, two- or three-word names like this should be treated as single entities. “General” only has meaning as a proper noun when it occurs in a phrase like “General Assembly” or “Attorney General Elliot Spitzer.” And a name with a title like that is treated as multiple NNPs, as in this example from the corpus:

```
[ Mayor-nominate/NNP Ivan/NNP Allen/NNP Jr/NNP ]
```

Doing a `grep` search for proper nouns returns a list of individual words. The regular expression `'([A-Z][a-z]+/NNP )+'` identifies entire compound proper nouns. Here is an excerpt from the output of a `grep` command using this regular expression:

```
American/NNP Army/NNP
Communist/NNP
Danzig/NNP
West/NNP German/NNP
General/NNP Clay/NNP
West/NNP Berlin/NNP
President/NNP Kennedy/NNP
Prairie/NNP National/NNP Park/NNP Thousands/NNP
American/NNP
```

The weakness with this output, of course, is that it presents difficulties in grouping different names for the same referent together—analogous to lemmatization. For example, If “President Kennedy” is mentioned in the first paragraph of a news story, he may be simply “Kennedy” or “the President” in the rest of the story.

There are errors and inconsistencies as well. For example, the word “States” in “United States” is sometimes tagged as a singular (NNP) and sometimes as a plural (NNPS) proper noun. And in one case, the proper name “Eisenhower” is identified as a comparative adjective.

Methodology

The data in the Brown corpus is made up of many small files. I order to use the data I needed efficiently, I first needed to concatenate the files that made up the two subgenres I needed. A `cat` command did this. The command

```
cat CB* > /export/home/jjd59/FINAL/editorial.POS
```

collected all the files that made up the Press: Editorial portion of the Brown Corpus. The files were named `CB01.POS`, `CB02.POS`, and so forth. The Press: Reportage files were concatenated in the same way.

Several new files were created from each of these concatenated files, based on parts of speech. For this project, proper nouns, adjectives, and adverbs were important. Using the `egrep` command in UNIX, I first extracted all lines that included the part of speech I was interested in and sent the output to a new file. Then I used `egrep` on that file to create a new file with just the words themselves in it. I kept the first `egrep` file in order to have context for the terms if I needed it. The resulting files are the ones that were compared to find the linguistic differences between the two types of newspaper content. Here are the commands used to create the "adjective" files for the editorial content.

```
> cat editorial.POS | egrep '(RB|RBR|RBS)' > adved1.txt

> cat adjed1.txt | egrep -o '([A-Za-z]+/JJ[A-Z]* )+' >
      adjed2.txt
```

The final step in UNIX is to use the `sort` and `uniq` commands to get listings of the proper nouns, adjectives, and adverbs, ranked by frequency.

Since the analysis of the data produced by the previous steps involves human interpretation, UNIX is of little help. So I opened the text files in Microsoft Word and did final manipulation of the data in that application and in Microsoft Excel.

Results

This study looked for potential differences in three areas: proper nouns, adjectives, and adverbs, and the two types of writing did indeed show distinctions in these areas.

The proper nouns, of course, are names of the people, places, and organizations discussed. The corpus is from the height of the Cold War, when the United States and the Soviet Union were rivals for power in the international sphere, and there was a constant concern that tensions could boil over into a "hot war." While actual newsworthy events did not take up much space in the reportage, the Cold War was certainly on the minds of the editorial writers. Nikita Khrushchev was the leader of the Soviet Union at the time. "Khrushchev" is the second most common term in the editorial content, but barely makes the top 50 in the rest of the paper. In fact, much of the editorial content seemed to be about Cold War issues. The top 20 terms included the following:

Khrushchev

Soviet

Communist

China

Berlin

West Berlin

Moscow

Russia

None of these terms occurred in the top 20 list for reportage, and only three of them (*Moscow*, *Communist,* and *Khrushchev*) appeared in the top 50. The reportage addresses more mundane matters: the top 20 terms include several days of the week and three athletes.

In analyzing adjectives and adverbs, I looked for words that might imply a value judgment, something less appropriate for news reporting. Because semantic judgments can be difficult without context, I chose to be conservative in deciding which words would be considered judgmental. The distinctions were much less clear than with the proper nouns.

As with the proper nouns, I listed the 50 most common words in each corpus for each part of speech, adjective and adverb. The list of adjectives with emotional or judgmental connotations was almost the same for the two groups. The editorial text included *great*, *least*, and *real*; the reportage included *great*, *best*, and *real*, but lower down in the list.

There was also little distinction between reportage and editorial content in the commonest adverbs. In the top 50 the editorials used *too*, *never*, *merely*, and *always*. The reportage used *most*, *never*, and *too*. Not very powerful words, even if they are superlatives. And not very many of them. This may be because these common words so dominate the text. Much further down the list in the editorial material are words used only once or twice like *grossly*, *courageously*, *sycophantically*, *grotesquely*, *fantastically*. Reportage contains a few words like *pitifully*, but not the more pungent ones found in the editorials.

So for proper nouns, the difference between editorial and reportage was the topics discussed. With adjectives and adverbs, the difference was in writing style.

Potential Future Research

Obviously, more sophisticated tools could make the more complex comparisons I considered at first, like whether editorial content uses more relative clauses or longer sentences.

These results were based on samples from multiple publications. The corpora are compilations of excerpts of about 2,000 words each from about two dozen newspapers and a few news magazines. This approach means that the results should be representative of American journalism, circa 1961, as a whole. It may be possible to use this data as a baseline for future studies to determine if an individual newspaper might have a bias. If the characteristics of editorial writing as identified by this study appear in the reportage sections, the publication may have an unspoken agenda to influence public opinion in the way it reports the news.

In any future research, particular attention needs to be paid to the semantic analysis. One unavoidable aspect of the material in the corpus is that it is all from calendar year 1961. While the current events covered in the newspapers will be broadly the same as in current newspapers—international tensions, crime, sports scores, the stock market, and so forth—the details, the players will be different. Countries have new leaders, of course, but there are also countries today that weren't on the map in 1961. The Dow Jones Industrial Average is still calculated every day, but six of its 30 component companies in 1961 are no longer in business. For this reason, a thorough analysis should include as background knowledge the events and newsmakers who were current in that year, and their significance in American history or culture.