

A Corpus-Based Model for the Work of Editors

Joseph E. Richardson
jerich@mstarmetro.net
March 13, 2008

Corporate Communications

- Products
- Originators and audiences
- Review and production processes
- Pragmatic function

Mediating Discourse

Approaches to usage and style prescription

- Intuition
- Style guides and usage books
- Example-based research

Available Corpora

- CR (1951–): 208,180 records (8.7 million words, with about 131,000 added each year)
- Magazines (1971–): 905,267 records (about 45 million words, with 1.1 million words added each year)
- Curriculum (mid-1980s): 135,258 records (about 2.1 million words)
- Handbooks: 13,474 records (358,140 words)
- Style Guide: 2,187 records (37,811 words)
- Three others: 54,024 records (1.6 million words)

Methodology

- Initial survey by e-mail
- Paper questionnaire several months later
- Follow-up request for additional information (by e-mail)
- Some interviews for clarification

Results

Received responses from 24/24 staff members

- Most editors use the corpora regularly (20).
 - Source checking: 17/20
 - Content search: 7/20
 - Style and usage examples: 20/20
 - Examples of issues addressed in the style guide: 13
 - Examples of issues not addressed in the style guide: 18
 - Examples of grammatical structure: 6
 - Examples of spelling unique to the organization: 14
 - Examples of particular usage: 16

Results: frequency

- Seldom: 3
- Monthly: 11
- Weekly: 5
- Daily: 1

Results: duration

- **Most searches are short**
 - 0–10 minutes: 17
 - 10–20 minutes: 4

Results: corpus preference

Conference Report (14/18)

- More authoritative (our product); carefully edited, proofread, reviewed
- Updated frequently
- Size of sample; reflects historical usage

Magazines (3/18)

- Largest sample
- Show a variety of styles
- Continually updated
- Perception that it is not as carefully edited

Results: corpus preference

Style Guide

- Not current or accessible

Results: typical search

- Typical search is to key in a word or phrase.
- How we have handled a particular feature (title, citation, etc.) in the past.
 - Examples of spelling and capitalization unique to the organization
 - Forebear, forbear, forebearers
 - steadfast or stedfast
 - OK vs. okay
 - Preposition use: “testify to” vs. “testify of”
 - Associative networks in language: i.e., *novel* functioning as an adjective

Frustrations

Context: in a production environment, search needs to be quick and easy

- **Difficulty restricting search**
 - Use of stop words prevents disambiguation
 - Certain characters (hyphens, apostrophes, and so on) have special search function
- **Ignorance of available search tools**
- **Query requirements**
- **Difficult navigation features**

Observations

Even without training in a corpus approach, most editors seem to prefer examples to rules. Editors seem naturally to use corporate databases for example-oriented research when they are aware such resources exist.

Observations

One editor observed that even when the answer to a question is available in the style guide or in Chicago, he often still goes to the databases to answer the question rather than to the style guides.

Observations

- Few editors want simple POS tagging by itself, but most would use it in addition to other tools.
- Editors could use training on how to make the most of corporate resources.
- The question remains whether decision makers in corporations will want to fund the extra time, effort, and resources to obtain software, tag for POS, and train writers and editors.

Observations

Reasons for POS tagging:

- Speed
- Specificity
- Coverage

Issues for POS tagging:

- How frequently is the corpus updated?
- Who created the text?

Observations

- It would be helpful to have a function built into our existing tools to enable us to manipulate the presentation of concordance lines, with ability to focus on key nodes.
- In corpora pertaining to corporations, many questions relating to representativeness, content, and size are easily resolved.
- Uses of corpora are rudimentary right now, but corpus research provides an opportunity for more detailed linguistic analysis, including content analysis.

Recommendations

- Create a single-source repository of corporate text with POS tags.
- Develop tools that overlay POS tagging on existing XML tagging structure and that enable easy annotations.

Recommendations

- Make corpus tools easy to apply in a production context, and promote their use.
- Teach editing and technical writing students to find, create, and use or exploit corpora in a production context.