

Discipline- and Genre-Specific Language Corpus Analysis— A Handy Tool for Clarifying Language Usage

Mary Ellen Kerans – Ailish Maher
Freelance translators & language editors, Barcelona, Spain

Problems—doubts from mixed language varieties, clashing idioms

Doubts about specialist word usage and combinations arise when we handle manuscripts outside the scope of our own reading and writing.

Solution—planned language sampling with a corpus

Use a 'target language corpus' of exemplary texts for guidance. A specialist corpus samples major topics and genres in a discipline, to reflect discourse community usage. Mine the corpus with a 'concordancer'.

Examples of problems & data-driven decisions: The tools used in the examples were *AntConc* and *WebCorp* (see description below). The corpus for *AntConc* was a 383,721-word sampling of peer-reviewed articles (pulmonology, thoracic surgery, anesthesiology) by native speakers of English, plus a few university or scientific society website tutorials. *AntConc* takes seconds to run. *WebCorp* can take a few minutes.

Simple problem 1, usual word order: *Right upper lobe* sounds strange to a lightly experienced author's editor's ear—native English (E1) speakers normally say *upper right corner*—so why is RUL appearing as the abbreviation in so many articles consulted?

Data-driven decision: Clearly, use *right upper*. The simplest *AntConc* output—key word in context (KWIC)—answers the question within seconds. Furthermore, hit 15 suggests the ordering is systematic for anatomy. The discipline- and genre-naive English-native-speaker (E1) editor obtained an answer and learned something new within seconds.

Full KWIC output for Problem 1:

File	Global Settings	Tool Preferences	About	Number of Concordance Hits
1	1	1	1	15
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5
6	6	6	6	6
7	7	7	7	7
8	8	8	8	8
9	9	9	9	9
10	10	10	10	10
11	11	11	11	11
12	12	12	12	12
13	13	13	13	13
14	14	14	14	14
15	15	15	15	15

Simple problem 2, register: Should 'Spirometry was done by personnel who had received training...' be edited to 'Spirometry was performed...' or 'Spirometry was carried out'?

Data-driven decision: No, don't overedit. At first it might look as if there is a tendency toward more American English use of *done*—hits 20–38 are from American journals. But don't jump to conclusions, as the writer of *done* was British and the editor American. Furthermore, this corpus is too small to settle such questions and wasn't designed to do so. Text PR AJRCCM.3 is a long consensus statement with sections by different authors, but a single editor seems to have imposed his/her idiolect. The first 19 hits (not shown) are from a continuous mixed-journal corpus of peer reviewed articles; it could be checked for patterns, but in fact the conjecture is unimportant to the immediate editing job. **Corpus analysis can help us cut down on overediting and the imposition of one editor's idiolect over others.**

Excerpt from KWIC output for Problem 2:

File	Global Settings	Tool Preferences	About	Number of Concordance Hits
1	1	1	1	38
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5
6	6	6	6	6
7	7	7	7	7
8	8	8	8	8
9	9	9	9	9
10	10	10	10	10
11	11	11	11	11
12	12	12	12	12
13	13	13	13	13
14	14	14	14	14
15	15	15	15	15

Less simple, problem 3: A translator has written *antiplatelet treatment* for *tratamiento antiagregante*. The editor vaguely thinks *therapy* sounds more familiar, but is a change justified?

Step 1: KWIC outputs from both *AntConc* and *WebCorp* (screenshots 1, 2, 3, 4) suggest that drug names and classes collocate with both words, and that both collocate with *antiplatelet*. A tentative conclusion is that either word can be chosen. (The screens shown are only excerpts.)

Step 2: Two *AntConc* tools—Word Cluster and Collocates—can clarify how a word combines with other words if that information isn't immediately apparent from the KWIC output. Screenshots 5 and 6 show an excerpt from the Word Cluster outputs for *therapy* and *treatment*.

Data-driven decisions: 1) Use *antiplatelet therapy*—it's the combination more readers will expect. 2) There seems to be a tendency to use *therapy* more often than *treatment* with the names of substances or interventions. 3) Words beginning with *anti-* collocate more often with *therapy*. 4) More adjectives, on the other hand, are found with *treatment*.

File	Global Settings	Tool Preferences	About	Number of Concordance Hits
1	1	1	1	1
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5
6	6	6	6	6
7	7	7	7	7
8	8	8	8	8
9	9	9	9	9
10	10	10	10	10
11	11	11	11	11
12	12	12	12	12
13	13	13	13	13
14	14	14	14	14
15	15	15	15	15

File	Global Settings	Tool Preferences	About	Number of Concordance Hits
1	1	1	1	1
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5
6	6	6	6	6
7	7	7	7	7
8	8	8	8	8
9	9	9	9	9
10	10	10	10	10
11	11	11	11	11
12	12	12	12	12
13	13	13	13	13
14	14	14	14	14
15	15	15	15	15

File	Global Settings	Tool Preferences	About	Number of Concordance Hits
1	1	1	1	1
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5
6	6	6	6	6
7	7	7	7	7
8	8	8	8	8
9	9	9	9	9
10	10	10	10	10
11	11	11	11	11
12	12	12	12	12
13	13	13	13	13
14	14	14	14	14
15	15	15	15	15

File	Global Settings	Tool Preferences	About	Number of Concordance Hits
1	1	1	1	1
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5
6	6	6	6	6
7	7	7	7	7
8	8	8	8	8
9	9	9	9	9
10	10	10	10	10
11	11	11	11	11
12	12	12	12	12
13	13	13	13	13
14	14	14	14	14
15	15	15	15	15

File	Global Settings	Tool Preferences	About	Number of Concordance Hits
1	1	1	1	1
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5
6	6	6	6	6
7	7	7	7	7
8	8	8	8	8
9	9	9	9	9
10	10	10	10	10
11	11	11	11	11
12	12	12	12	12
13	13	13	13	13
14	14	14	14	14
15	15	15	15	15

File	Global Settings	Tool Preferences	About	Number of Concordance Hits
1	1	1	1	1
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5
6	6	6	6	6
7	7	7	7	7
8	8	8	8	8
9	9	9	9	9
10	10	10	10	10
11	11	11	11	11
12	12	12	12	12
13	13	13	13	13
14	14	14	14	14
15	15	15	15	15

Setting Up a Target-Genre Corpus — Steps Toward a Data-Driven Approach to Language Editing Decisions

Main principles for corpus building

- 1) Define the disciplinary scope, eg, respiratory medicine & thoracic surgery; antennas & signal processing
- 2) Define the target language scope, eg, language a) in a set of peer-reviewed journals since 2000; b) in certain genres (case reports, IMRD, proof-of-concept); c) by E1 speakers or generated in E1 environments
- 3) Gather texts to sample the discipline's main topics.
- 4) Store texts in 2 ways: a) intact (html or pdf files) for analysis of major features, and b) as txt files, cleaned of programming artifacts, graphics, references, to load into a concordancing program.

Obtain the freeware concordancer *AntConc*

Fast, easy to learn to use. Download for Windows or Linux from developer Laurence Anthony's website: <http://www.antlab.sci.waseda.ac.jp/>

Other corpora and tools to use and watch for (as of spring 2006)

- **Corpus of Professional English (CPE)**, searching by broad disciplines; fee, predicted for December 2006; project of the Professional English Research Consortium (PERC): <http://www.perc21.org/>
- **WebCorp**. The WWW as a corpus. Crude subspecialty searching (eg, science, health, sports, etc): <http://www.webcorp.org.uk/>
- **British National Corpus**, 100m words, trial license followed by small fee; no subspecialty searching: <http://www.natcorp.ox.ac.uk/>
- **Bank of English**, fee for access to 200m words, search for Br vs Am: <http://www.titania.bham.ac.uk/dccs/>; free sampler of 56m words at <http://www.collins.co.uk/Corpus/CorpusSearch.aspx>.

Further Reading

Anthony, L. (2005). *AntConc: Design and Development of a Freeware Corpus Analysis Toolkit for the Technical Writing Classroom*. In G Hayhoe, ed. 2005 *IEEE International Professional Communication Conference*. Piscataway, NJ, USA: IEEE, pp.729-37.

Anthony, L. (forthcoming). Developing a Freeware, Multipatform Corpus Analysis Toolkit for the Technical Writing Classroom. *IEEE Transactions of Professional Communication*.

Hunston S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Kerans ME. (Forthcoming, 2006) *Grammarians or linguists? On using language corpus data to guide usage. The Write Staff*.

Maher, A. (2006) *WebCorp as a translation resource. Identifying candidate terminology from concordance and collocate data: a practical example. Part 1. Cadaceus*, pp. 18-2. Available at http://www.atadivisions.org/MD/Caduceus_2006Spring_2.pdf

Maher, A. (Forthcoming, Summer 2006) *WebCorp as a translation resource. Identifying candidate terminology from concordance and collocate data: a practical example. Part 2. Cadaceus*.