

Grammarians or linguists? On using language corpus data to guide usage

by Mary Ellen Kerans

The grammarian and the linguist talk about the same objects—the rule-driven patterns that sounds and symbols make—but their approaches are very different. I'll argue that the attitude of one type of linguist—the 'corpus linguist'—is more relevant for technical writers, editors and translators because it helps us grasp subtle usage differences between disciplines efficiently. Embarrassingly easy to begin applying, a linguist's approach and tools enable us to handle language confidently but respectfully. The grammarian's approach is limited, authoritarian, and often idiosyncratic. Both can produce appropriate texts—but a linguist is more versatile and quicker at finding solutions to new problems.

Opposite poles

Introductory linguistics courses distinguish the two simply: the grammarian is prescriptive while the linguist is descriptive. The grammarian imposes structure, whereas the linguist looks for the system that has evolved and is expected to change. Proof of that concept is given to budding linguists in stories of the earliest English grammars, which were little more than imposed Latin categories at a time when English variation was great and confusing to the newly, barely literate of the generations following the introduction of print. Today the paradigmatic grammarian is a remembered school teacher: bespectacled and judgmental. The paradigmatic linguist? The missionary-anthropologist writing up a dying tribe's oral language, maybe: Indiana Jones with an interest in tgmemes and the like.

That distinction, like all dichotomies, is simplistic. David Crystal [1] tells of more benign early interaction between Latin and English: 'Participial constructions became extremely common and added greatly to the length of sentences.... There was conscious experimentation with new grammatical patterns.... By the 17th century, highly sophisticated and carefully crafted sentences, following a variety of Latin models, were commonplace...' (p. 70). And the new grammarians of the late 20th century were certainly benign and helpful in their highly descriptive attitude. Corpus linguists Randolph Quirk and Sidney Greenbaum broke ground in the early 1970s with grammars welcomed on both sides of the Atlantic and published in various forms. (For instance, see the American edition of 1975 [2].)

Still, a distinction remains clear. However descriptive a grammar might be, we refer to it saying, 'Quirk and Greenbaum say....' Grammar books are meant to provide

authority, guidance: we consult them. I see polarization separating prescriptive and descriptive camps whenever I'm privy to practical discussions of language—in publishing houses, on e-lists, in teacher's workrooms, at translators' and editors' meetings. Grammarians worry that a descriptive approach is too democratic, saying that it will have us following horrendous examples. Where they think bad examples come from varies. I've heard declining English standards blamed on 'EuroEnglish', on users from the western side of the Atlantic, or on the undereducated young. It seems politically incorrect to ascribe them to the unwashed masses these days, but non-native users of English—even highly educated post-colonial ones—take hard knocks.

Why take up with linguists?

The linguist's statements, on the other hand, are driven by data (samples collected according to specifications) and are couched in non-judgmental terms. Observations are of usage in characterized text samples—a 'corpus', which might hold as few as 200,000 words or over 100 million like the British National Corpus [3]. A linguist reads a corpus the same way Luther, Knox and Calvin's followers read the first printed vernacular Bibles—in a community of watchful peers but in the expectation of personal revelation.

Such a cocky attitude is heaven-sent for people like me. By 'like me' I mean many readers of this paper—those who need to satisfy high expectations in varieties of language that may not come naturally. We are writers of other people's ideas—whether journalists, ghosts, co-authors, translators, editors or anthropologists. We may be editors and translators in specialties we were not educated in. We might work in languages other than the tongues of our mothers or teachers. Some of us are native speakers who have long lived abroad and acquired a hesitant familiarity with varieties and registers we didn't grow up with, concerned our learned usage may not be quite right.

Example of how a corpus linguist approaches problems

Acquiring the approach is easy—but as with riding a bicycle, skill comes with practice. I'll now show how a problem can be solved through corpus consultation. I'll then name some of the basics one needs to know to create a specialized corpus like the one used in the example.

Not long ago, *TWS* editor Elise Langdon-Neuner posted a query on the e-mail listserve of the European Association

Grammarians or linguists?

of Science Editors (EASE-Forum) [4]. A copyeditor had changed ‘Based on these findings, we depleted T cells from spleen cells...’ to ‘From these findings, we depleted...’ Elise wondered what ‘the linguists’ thought of an apparent ban (later confirmed by the journal) on beginning a sentence with *based on*.

Figure 1: A KWIC display generated by AntConc, a freeware concordancer that helps a linguist—or any practical wordsmith—analyze usage patterns. With the view scrolled to the right, it’s possible to see that based on is used in this 350,000-word corpus of peer-reviewed articles by native speakers of English to introduce phrases that function as context frames for the sentences that follow.

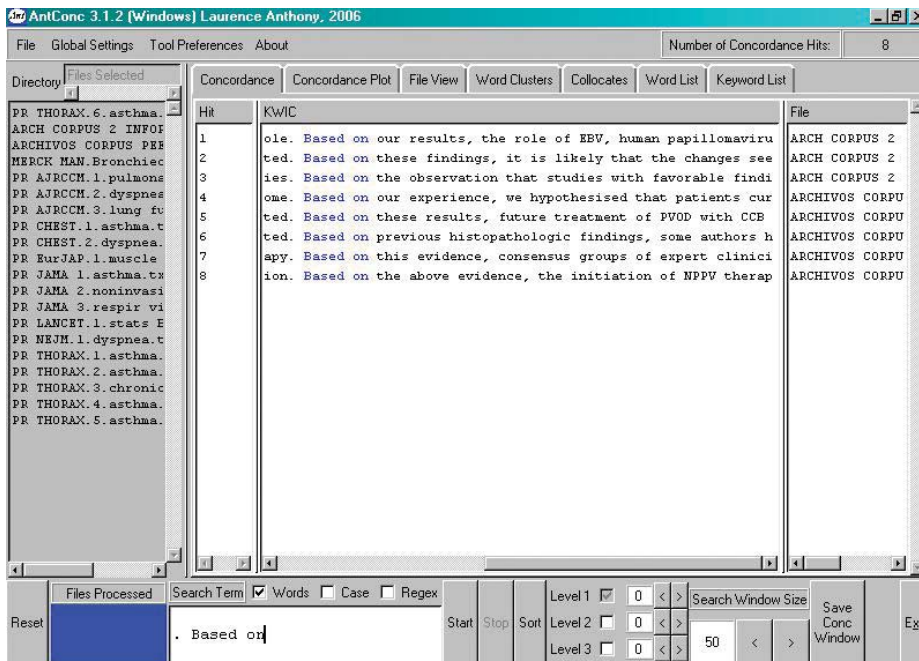
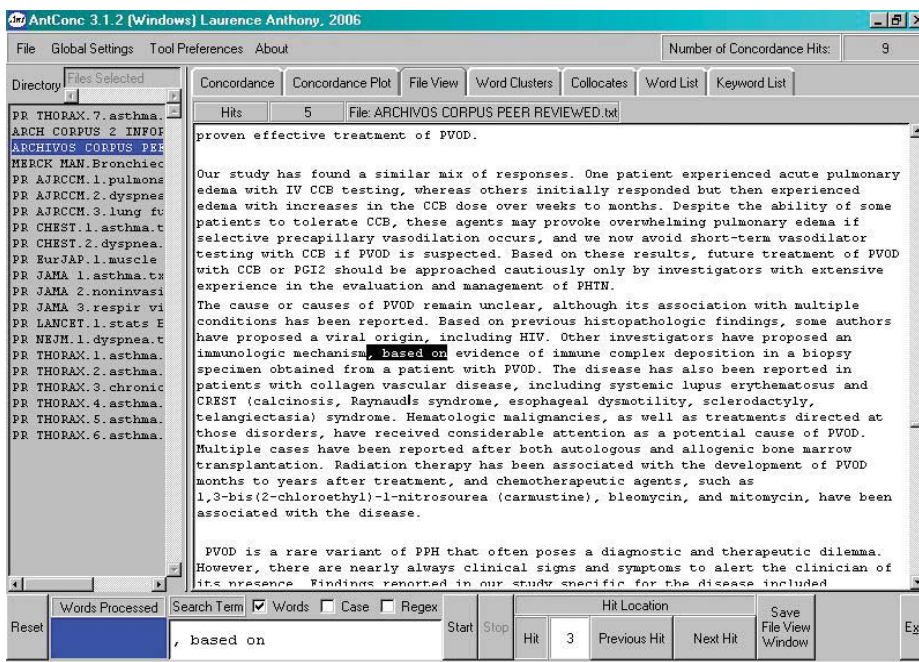


Figure 2: ‘File View’ display that reveals the adverbial nature of a *based on* hit that looked as if it might be adjectival in the shorter KWIC display. Clicking on any hit in an aligned KWIC display in the freeware AntConc concordancer toggles quickly to this view.



I participate in the Forum and remember my first reaction to the invitation. First of all, I thought, most linguists make it a point not to think too much about any text problem until they’ve looked at a corpus. So I entered *based on* into the search box of a concordancing program open on my desktop. Such a program—with texts already loaded—lets me see relevant patterns in seconds through a display called ‘key word in context’ (KWIC). (See Figure 1.) The loaded corpus contained articles in a discipline similar to that of the Forum query: papers from respected peer reviewed journals plus a few tutorials from teaching hos-

pital websites and similar sources. The corpus is designed to give me information on language usage and expectations in a ‘discourse community’. That term comes from ‘genre analysis’, a variety of corpus linguistics that looks at large rhetorical patterns as well as phrasal ones. Swales [5] defines a discourse community as a set of individuals with a common goal and means for member communication (information and feedback) that include specific genres, or text types, and lexis.

It was obvious from the first KWIC display of 136 hits (not shown) that many were in a verb phrase plus complement construction. They were irrelevant to the query, so I searched again with a period and *Based on* in the box. Figure 1 shows 8 good hits of uses exactly like the one in the EASE-Forum posting. For a small corpus (about 350,000 words), and for a specific construction and word, that frequency is high. Normally a consultation like that, taking seconds, would be enough to assure me that the author’s wording should be left alone. Had I been the journal’s copyeditor and insecure about participles because I’d been keeping company with grammarians, the KWIC display would have reassured me that the author’s discourse community would be as comfortable with the phrase as the author and I were.



>>> **Grammarians or linguists?**

Looking for a more complete answer for a posting to the EASE-Forum, though, I continued querying. I'd noticed several of the original 136 hits were preceded by commas—suggesting they might be end-of-sentence versions of the introductory participial phrase. So putting a comma plus *based on* into the search box, I retrieved 9 more hits. Of those, 4 were discarded because they introduced post-nominal adjectival modifiers (eg, 'An alternative approach, based on...'). Four hits were adverbial sentence endings similar to the introductory phrase. (If this adverbial use of *based on* falls at the beginning, it would introduce a 'context frame', as it's called by the newest linguist/grammarians who analyze 'systemic functional grammar' [SFG]. Falling at the end, it would be part of a 'rheme' or the rheme itself if it contains the kernel of new information focus. SFG analyzes units that wed structure to rhetorical function.) To complete the analysis, I needed more context to confirm that an end-sentence hit was also adverbial, not adjectival. Clicking on the word to open the 'file view' (Figure 2) confirmed both pattern and sentence adverbial function clearly.

Thirteen hits for an adverbial use of *based on* in an appropriate corpus strongly supported Elise's complaint against intrusive copyediting. I'd looked at 4 displays in about a minute and a half. Concluding my EASE-Forum reply, I said that the sentence was well supported in its original form and the copyeditor had improved nothing by implying that T cells were depleted from spleen cells from findings!

I'll concede that efficient small corpus consultation does not solve all problems. The approach was useful in a similar EASE-Forum query on the use of the suffixes *-ic* and *-ical*, but for another about multivariate versus multivariable analysis, it was relevant but insufficient. We needed statistical concepts to answer that question, as frequency and patterning were not the whole issue. So, while this approach is pivotal for one who is outside a target discourse community, it's no substitute for knowledge.

Do you really need a corpus?

Could I have found the answer in a grammar book? Both Quirk and Greenbaum [2] and Fowler [6] in his first edition discussed the matter of the 'attachment' of such participial sentence introducers. To check Quirk and Greenbaum I needed to know that the section called 'non-finite or verbless clauses' would be helpful. Those authors give sensible, open-minded statements about the 'attachment rule'. Fowler's advice, found under 'unattached participles', comes in his superior tone. Both books mention exceptions. Here are excerpts from the entries:

... Commonly, however, this 'attachment rule' is violated:

?Since leaving her, life has seemed empty

In this case, we would assume that the superordinate clause means 'Life has seemed empty *to me*.... Such 'unattached' ('pendant' or 'dangling') clauses are frowned on, however...

Note

[a] The attachment rule does not need to be observed with disjuncts:

Speaking candidly (S='I'), John is dishonest

[Quirk and Greenbaum, p. 329]

...[It] is to be remembered that there is a continual change going on by which certain participles or adjectives acquire the character of prepositions or adverbs, no longer needing the prop of a noun to cling to; we can say *Considering the circumstances you were justified, or Roughly speaking they are identical*.... The difficulty is to know when this development is complete.... In all such cases, it is best to put off recognition. A good example of what may prove to have been such a development caught in the act is the phrase *due to*. Every illiterate in the land is now treating *due to* as though *due to* had passed into an adverb not needing a noun to agree with, just as *owing*, in *owing to*, has actually done....

[Fowler, p. 675]

Perhaps it's personal, but those grammarians leave me feeling less confident than the foregoing linguistic approach did. Quirk and Greenbaum's question mark makes me feel I might be naughty to use such frowned-upon sentences and the negative term disjunct—exactly what *based on* is—might give me permission to dangle the introductory phrase bravely, or it might leave me still feeling sheepish. And how is one to know if a phrase has become an adverb like *owing to*, as Fowler says, or a disjunct to which the rule does not apply, as Quirk and Greenbaum say? There's an implied caste of people who know such things and one suspects that if one has to consult a book one might not be a member. If we keep consulting, we'll see that *The New Fowler's*... of Burchfield [7] spares us allusions to illiteracy but explicitly admonishes us not to use *based on* to introduce sentences; and where Fowler hedged by referring to expressions acquiring the character of adverbs (correctly I think) or prepositions, Burchfield confuses me by referring to 'marginal prepositions' for the same usage shift (p. 804) (incorrectly I think). It's no wonder that copyeditors are intimidated into changing to *on the basis of*, though to the linguist that looks no better founded.

Grammarians or linguists?

How to set up as a linguist

The main concept to grasp is that it is an attitude toward language and an approach to problem solving you're buying into. The approach is through the description of patterns—based on adequate sampling and a well-characterized corpus of exemplary language—with the support of member informants for complex questions as in anthropology. I'd characterize the attitude as one of respect for specific discourse communities. This leads to defining exemplary users of language worth emulating. A part of the definition of such communities I haven't mentioned yet is that membership changes and community survival depends on a threshold level of individuals with discursal expertise [5]. This is why corpus design is the key to the validity of observations and why the corpus must be well characterized. To borrow the grammarian's aura of expertise: any authority a linguist's appraisal might have rests as much on the appropriateness and adequacy of the corpus as on discernment.

Building a specialist corpus takes time, but how much depends on how fussy you are. My largest corpus, now at about 350,000 words, is fairly clean and I know exactly what's in it. A clean corpus is free of the sorts of artifacts that web pages insert, of images and tables that take up space but give little language information, and of references. It's also free of duplications, which plague corpus builders who collect haphazardly, job by job. Logging texts and cleaning off artifacts like html or other coding takes a bit of time, but that's no excuse for ruling out corpus creation. Even a quickly assembled 'dirty' one can give good service [8]. I know someone who works happily with a dirty corpus of a million words compiled in little time and thinks cleaning unnecessary. Still, I've found it's worthwhile to have a display that's rich and compact and find very dirty corpora more difficult to scan and draw inferences from. As for size, I started to get useful information from a highly specialized corpus at only 200,000 words.

How to build a corpus lies outside the scope of this article, but here are some principles for anyone who'd like to get started:

- Define the scope of the language you're interested in, state how you'll identify exemplary texts, and find a free electronic source.
- Store samples in 2 ways: as intact files and as text files. This is fair use if the corpus is for personal research or for sharing with a small group of colleagues or students.
- Name text ('notepad' files) and intact (html or pdf files, for instance) identically so their correspondence and content will be evident at a glance and log them so you know what you have. Use intact files to study rhetorical patterns and process text files in a concordancer to study frequency and collocations.

The concordancer I recommend is freeware—AntConc, available for Windows and Linux [9]. Developed by Laurence Anthony, currently at Waseda University in Tokyo, AntConc is simple and the two most basic displays (KWIC and file view) are intuitive. For more advanced functions, Anthony's help file is short and indeed helpful.

To know more, read Susan Hunston's [10] overview. It's no more technical than necessary.

The linguist's approach is useful, efficient and enjoyable. I am a wordsmith—not a proper applied linguist. But this attitude toward language has been a key to acquiring competence quickly when I've had to deal with new varieties or genres. It's also helped me avoid insecurity while working abroad in a language without an Academy, with competing regional varieties and awash with prejudices that grow out of overgeneralization from idiolect. Concepts from linguistics help me form sets, describe patterns and develop algorithms for applying them. Like statistical data processing, corpus analysis is an aid to inductive reasoning and can be used to generate or refine a hypothesis, falsify one, or simply confirm intuition.

Acknowledgments

Ailish Maher and Iain Patten gave helpful comments on an early draft.

Mary Ellen Kerans

Barcelona, Spain
mekerans@telefonica.net

Mary Ellen Kerans is a translator and editor. She is a co-developer, with Ailish Maher, of a workshop on corpus-guided translation—part of the activities program of the association Mediterranean Editors and Translators: <http://www.metmeetings.org/index.htm>.

References

1. Crystal D. The Cambridge Encyclopedia of the English Language, Second Edition. Cambridge: Cambridge University Press, 2003
2. Quirk R, Greenbaum S. A Concise Grammar of Contemporary English. New York: Harcourt Brace Jovanovich, Inc., 1975
3. British National Corpus. <http://www.natcorp.ox.ac.uk/>.
4. Langdon-Neuner E. EASE-Forum digest: October-December 2005. European Science Editing 2006;32(1):18–20
5. Swales J. Genre Analysis: English in academic and research settings. Cambridge: Cambridge University Press, 1990
6. Fowler H. A Dictionary of Modern English Usage. Oxford: Oxford University Press, 1926
7. Burchfield, R. The New Fowler's Modern English Usage. Oxford: Oxford University Press, 1998
8. Tribble, C. Improvising corpora for ELT: quick-and-dirty ways of developing corpora for language teaching. In: Lewandowska-Tomaszczyk B, Melia P (eds.) International Conference on Practical Application in Language Corpora, Łódź, Poland, 11–14 April, 1997. Proceedings, p. 106–17. Łódź: Łódź University Press, 1997
9. Anthony L. AntConc3.1.302. <http://www.antlab.sci.waseda.ac.jp/>
10. Hunston, S. Corpora in Applied Linguistics. Cambridge: Cambridge University Press, 2002

Word usage: Let the Web decide!

A useful site for finding the preferred spelling of a word, which of two words is more commonly used or whether a word is more often or not hyphenated is <http://www.spellweb.com>. Here you can enter your two alternatives, choose the search venue, either Google, Alexa or Yahoo, and SpellWeb will come up with the number of times each alternative appears in your selected venue and pronounce the winner.

langdoe@baxter.com