Web searches should supplement judgements, not supplant them

CARSON T. SCHÜTZE

The use of the World-Wide Web in linguistic research is increasing, and with good reason: it is the largest and most up-to-date linguistic corpus that has ever existed, it continues to grow and remain current, and it is (in limited ways) automatically searchable. The Web as a corpus can be useful to linguists in two kinds of ways that it is important to distinguish.

One kind are what I call idea-sparking uses. Here the Web plays the same role as a sentence you might happen to have noticed while reading the newspaper, listening to the radio, etc. You have simply increased your odds of “running into” an example relevant to your current research by using a search engine to pick your reading material. The Web provides examples that you might not otherwise have thought of, but having done so, it plays no further role in linguistic argumentation. It is as if the example had simply sprung to mind spontaneously, or dropped out of the sky and landed on your desk — in neither case do we conclude anything about the status of the example — its (degree of) grammaticality, its meaning, etc. — until further investigation is done. If this investigation leads nowhere, there will be no reason to report having found the example. If it leads to a fruitful discovery, there may still be no reason to report the particular example found on the Web: invented examples might serve your rhetorical purposes better. The crucial point is that in this scenario the examples found on the Web are not themselves used as evidence to support any claims about the language. You might have initiated a search because a question arose in your research — “I wonder if anyone can say X?”, but the search hit(s) are not taken to answer that question in any direct sense.

It is hard to imagine anyone objecting to a linguist using the Web in the way just discussed. Science does not generally have rules about how you are allowed to come up with your ideas, as long as you do not steal them from another scientist. (Hopefully you examine your search results in enough detail to know whether they came from a linguistics article.) The above idea-sparking scenario is to be contrasted with the evidentiary
scenario whereby Web data are used as evidence to support a particular claim about the language that they (apparently) exemplify. Broadly speaking, one can find two such kinds of uses: qualitative and quantitative. The qualitative claims typically involve (non)attestation as evidence for (im)possibility in the language; this is the case I will focus on. The quantitative claims generally presuppose the same assumptions about attestation and in addition interpret the relative frequencies of different forms or constructions vis-à-vis the theory, e.g., more frequent forms should violate fewer grammatical constraints. Independent of the nonobvious linking assumptions required to interpret corpus counts in this way, there are nontrivial technical challenges posed by the way that search engines extrapolate from smaller samples to estimate the number of hits in the entire Web corpus; these issues are beyond the scope of this note.

Concerning attestation in the qualitative evidentiary scenario, the main point I wish to make is that merely having found instances of a construction of interest should not be construed as evidence of anything ipso facto. There are many reasons why finding a sentence on the Web does not imply that even a single speaker of the language would produce or accept that sentence.

Among other possible explanations, the sentence could have wound up on the Web because ...

1) it was produced by a non-native speaker;
2) it was produced by a (native) speaker who intended to produce something else but made an error, because she was distracted, impaired, etc., or because she made a subsequent edit and did not check whether it fit in the sentence as a whole;
3) it was automatically generated, e.g. by translation software, voice recognition software, or optical character recognition, and the process was not perfect.

Furthermore, even if the sentence was genuinely produced, to be useful for linguistic theorizing one generally needs to know its intended meaning, which may be hard to discern with certainty from the Web context, just as it may be hard or impossible to determine who or what produced the sentence, whether it was proofread, etc., not to mention when and where it was produced.

Of course, problem 2), unlike the other two, is true also of traditional (written) corpora, because unlike standard elicitation or judgement-gathering situations, it is extremely difficult to ask follow-up questions of the source of an utterance found on the Web. And the number of typos on
the Web relative to the base rate of correct productions can be surprisingly high, as in the following recent Google™ hit counts:

“I beleave” 103,000
“I beleive” 2.76 million
“I believe” 184 million

These numbers raise another important point. Probably no one would try to use a lone example of some phenomenon from the Web as evidence; everyone expects that a real phenomenon should show up multiple times in a corpus as huge as the Web. But how many is enough? A priori, how could one know that 2.76 million hits was far too few to represent the dominant pattern for the above search? Once we are searching for larger structures it becomes almost impossible to estimate priors, i.e., how many hits we would expect for the target construction given the frequency of some related construction(s). And the problems do not end with finding multiple exemplars. It makes a difference to how we interpret them whether, say, 50 hits came from 50 different speakers or all from a single speaker (if we can even determine this).

Given all of these potential problems, most of which are obvious to anyone who has spent any time on the Web, it is informative to examine how Web data are actually being used in linguistic articles. A common pattern is exemplified by Bresnan & Nikitina (2003), whose descriptive goal is to show that the range of ditransitive verbs in English that can occur in the prepositional dative frame as well as the double object frame is much larger than generative linguists have claimed. To this end the authors quote a handful of examples from the Web, such as (1):

(1) This life-sized prop will give the creeps to just about anyone!

Tellingly, however, they follow the examples with the following remark: “We judge the web examples given above to be grammatically possible. [Footnote: Thus we do not classify our usage data given above with the sporadic adult errors.]”. In other words, they acknowledge that merely finding these sentences on the Web could have been consistent with them being production errors, and that their own intuitions played a role in concluding that the sentences are genuinely grammatical. Thus, the use of Web examples by Bresnan & Nikitina (2003) is of the idea-sparking type discussed earlier, at least in part, and not purely evidentiary. The examples inspired the authors to collect judgements data (from themselves) on sentences of a kind that had not been considered in this literature before, and the evidence they present consists of both (selected) search results and judgements. It appears that the mere existence of the
Web examples might not have been considered conclusive by these authors; in my view this would have been the appropriate stance to take.

The same is true for Sampson (2007), who argues vigorously against any use of “intuitions” in the scientific study of language, but then after citing eight examples from a Google™ search of a configuration that had been claimed to be impossible in earlier linguistic literature, he feels the need to add, “They make sense and feel normal enough to me (and, I suspect, to my present readers)” (Sampson 2007: 17). He thereby reports not only his own linguistic intuitions but also his predictions about other people’s. (This criticism is also raised by Meurers (2007).) Given his theoretical stance, it seems unlikely that he would have done so if he believed that the search data could stand as evidence on their own.

Contrast these two cases with the extended discussion by Manning (2003) of his chance encounter with the sentence (2) in a novel:

(2) By the time their son was born, though, Honus Whiting was beginning to understand and privately share his wife’s opinion, as least as it pertained to Empire Falls.

(Manning 2003: 292–294)

Manning wonders whether the sequence as least as was more than a typo, since his idiolect (like mine) would have required at least as in this context. He decides to bring the full force of corpus linguistics to bear on the question:

A search of 1994 New York Times newswire yields no examples — is this just because they have better copy editors? — but then I find four examples … in 1995 New York Times newswire, and two more examples from 1996. It already appears less likely that this was just a typo. A search on the Web (which with every passing year becomes a better medium for empirical linguistic research) yields hundreds of further examples … Finally, I find examples with at least as in a neighboring sentence …, showing intraspeaker variation. While much less common than at least as (perhaps about 175 times less common, based on this small study), as least as seems to have robust support … [many] of these additional examples use as least as in conjunction with an as Adj as construction, which perhaps provides a pathway for the development of this apparently new form. (Manning 2003: 293)

That is, solely on the basis of these searches (chiefly the Web search), Manning not only concludes that as least as is a genuine phenomenon of English (for some speakers), but that it is in (free?) variation with at least as for individual speakers, and he formulates a hypothesis about
Web searches should supplement judgements, not supplant them

how it could have developed historically. (His conclusions are endorsed by Penke & Rosenbach (2004).)

At no point in this study does Manning report having asked any English speakers (other than himself) whether they can actually say *as least as*, whether they consider it well-formed, whether they notice anything wrong when he reads sentence (2) out loud to them, or any other question that could rule out the possibility that these attestations, though not scarce, might indeed be typos. Evidently it is the sheer number of hits that compels Manning to reject this possibility. But this is where the issue of base rates becomes highly relevant. If these were typos, they would be easy to explain: a slip from *at least as* to *as least as* would be an anticipation of the *s* triggered by the fact that at least two other *a—s* sequences need to be typed imminently in the same phrase, including one with the identical context (#_#), and sometimes three, in the frequent situation Manning observes where the sequence *Adj as follows as least as* (e.g., “*as least as important as …*”). This anticipatory substitution is thus a highly plausible typo. What we would like to know is what percentage of the time we would expect it to occur, given the kinds of typing situations in which all the instances of *as lat least as* on the Web were produced. If the answer is at least 0.6% of the time, then an appeal to any other (e.g., linguistic) explanation for the presence of these phrases would be unjustified.

Another base rate that could easily be extracted would be the relative frequency with which *as least as* versus *at least as* occur in a neighboring sentence once a first occurrence of *as least as* has been found. Then we would be in a position to evaluate Manning’s observation that some number of such cases involve apparent “intraspeaker variation,” i.e., inconsistency. My suspicion, based on Manning’s report of the 175:1 frequency ratio, is that most instances of the *as* variant that occur near another instance of the same construction will be of the “inconsistent” type. That is, it is probably relatively rare to find the *as* variant used consistently within a passage. To me this state of affairs would be further support for its status as a typo. Oddly, Manning cites such intraspeaker variation in support of the grammatical status of the *as* variant, along-side its attestation in English dialects from different parts of the world. Both observations are obviously fully consistent with my slip-of-the-keyboard counter-hypothesis.

My view is thus that evidentiary uses of Web data require checking those data against the intuitions of live human speakers of the language. In case such verification fails, consideration should be given to how the exemplars might have arisen *other than* as genuine instances of use of the language in question, and to how their frequency compares to relevant base rates.
References


cschutze@ucla.edu
University of California, Los Angeles