Internet Searches as a Tool in Syntactic Research

Motivation

Collins and Postal (2012, 2014) contain lots of data obtained from Internet searches. Such data is also cited extensively in chapters 2 and 3 of my forthcoming monograph "Principles of Argument Structure: A Merge Based Approach". In this blog post, I offer a few guidelines on using such data in syntactic research.

Internet searches have turned out to be a revolutionary tool in syntactic research. Here are a few reasons why. First, if the domain you are looking at is controversial, Internet searches afford a way of finding non-elicited English examples that might help resolve the issue. For example, chapter 3 of my monograph claims that secondary predicates can modify the implicit argument in the passive. That is controversial since various authors have claimed that secondary predicates cannot do so. However, it is quite easy to find relevant examples on the Internet which I find completely acceptable.

More generally, Internet searches are a tool that can give the syntactician confidence in their claimed empirical results. Suppose that I propose a particular generalization which has not been investigated before. If I am able to easily find examples on the Internet conforming to the generalization (and crucially, I find those examples to be acceptable) then I will have increased confidence in my generalization.

Second, if you are looking at a relatively unexplored data domain, Internet searches can help to fill out the range of combinatorial possibilities. For example, in chapter two of my monograph, such searches helped me to figure out the range of possible phi-feature values of the implicit argument in the short passive. Once you have dissected the combinatorial nature of any particular problem, you can run searches on all the various possibilities, greatly expanding your knowledge of an empirical domain in a short period of time. Furthermore, tracking down and documenting the combinatorial possibilities will often lead to surprising discoveries.

In this blog post, I outline a methodology that takes Internet searches to be a tool used by the syntactician in syntactic research.

Searching

The basic technique is to search for phrase types that are being investigated. For example, in chapter 2, I investigate the use of anaphoric expressions like *on my own* in short passives. So I would search the following, where the quotes ensure that a string, not just a set of words, is the target:

(1) Google: "was done on my own"

In this example, I included "was done" to make sure that the search includes a passive participle, not an active participle. However, I have left out the subject, because examples would be relevant

no matter what their subject is. Some of the hits may be completely irrelevant so you may have to scroll through several screens to find good examples.

One of the hits of the search in (1) is:

(2) None of this was done on my own. To me success always takes collaboration.

This example is useful to my research. In my monograph, my hypothesis is that the implicit argument in the passive is syntactically active. (2) supports that hypothesis, since the antecedent of *my* is the implicit argument (the doer). (2) is also helpful because the second sentence clarifies the meaning of the first.

Follow-up Searches

Follow-up searches based on (1) can be done, yielding further information. For example, to investigate the phi-features of the implicit argument, *my* in (1) can be replaced by a whole range of possessive pronouns: *yours*, *his*, *hers*, *ours*, *theirs*. Alternatively, the main verb could be changed from *done* to *written*, *created, built*. Lastly, to vary the syntactic context, the copula could be changed from *was* to *were*, *is*, *are*, *to be, will be*, *been*. Just these choices would yield 5x3x6 = 90 additional searches, any of which could yield interesting example sentences.

See Collins 2018 for the use of the * operator in Google searches.

Verify with Target

I use the Internet as a tool for looking into the properties of a particular dialect, usually my own I-language (the target of investigation). In doing Internet searches, I am looking for data that I myself find acceptable. For example, after doing the search in (1), I found (2). Crucially, (2) is acceptable for me as a native speaker of English. If I find a relevant example, and I judge the sentence as acceptable, then I save it to a list for future use in a paper or book. If I judge the sentence as unacceptable, I put it aside for further consideration (see next section).

If you are looking for a construction that is not from your own dialect, you need to find a native speaker of the relevant dialect to verify it with.

What to do with Unacceptable Sentences?

Suppose that you do a search and find some a sentence S that you yourself find unacceptable. Furthermore, S is relevant to the theoretical issues you are looking into. What then? There are two routes to take at this point. First, you can do further searches to make sure that S was not just an error of some kind. If there are many such examples, and they look like they were produced by native speakers of English, then that is evidence that the S is acceptable for some native speaker. Second, you can try to find a native speaker of the dialect in question. Go to Facebook or Twitter, and post a query about the sentence you found. If there are native speakers, you may be able to find them in that way. But any serious investigation of S will have to involve native speakers who find S acceptable.

Quality Control

Of course, there is no guarantee that the data on the Internet will be of high quality for syntactic research. It is completely uncontrolled, so caution needs to be exercised. I propose the following guidelines.

Control 1:
Compare the string searched to actual hits. For example, for (1) above, the hits included the example "Today's work was done on my own fingernails…". Although the hit contains the search string, it is not the kind of example I am interested in (because it does not use the expression *on my own*).

Control 2:
Immediately after finding an example in the search result list, you should check to see that the URL is active and that the sentence actually appears on the accompanying website.

Control 3:
You should examine the sentence to make sure that it does not represent any of the following: (a) a clear grammatical error, (b) humorous writing (which plays with language), (c) poetic license (again playing with language), (d) an AI generated text (not produced by a human), (e) a Google translation (again not produced by a human).

Control 4:
A related concern is whether the sentence is being used as an example in a linguistics paper online. If so, it might be elicited data constructed by a linguist, and not non-elicted Internet data.

Control 5:
While at the website, you can try to see if there is anything that might indicate that the example was created by non-native speakers of English. If there are red flags, you should discard the example. For example, does the text in the website generally show signs of being written by a non-native speaker?

Control 6:
You should look at the context of the hit, in the preceding and following text. Such context might shed light on the interpretation of the examples, and might be useful to include in your work. For example, in (2), the second sentence explains the first sentence.

Number of Hits

The number of hits of a particular search can be significant. If I search for S and I find scores of high-quality hits (satisfying all the controls above), that is significant. Furthermore, if I search for S, and do not find any hits (or only one or two suspicious hits), that is also significant. But other than this basic dichotomy (many versus zero/few) it is not recommended to rely on actual counts. For example, what would I make of the fact that when I search for S1 I find 67 hits, but when I search for S2 I find 112 hits?

Here is the way we put such frequency results in Collins and Postal 2012:

(pg. 1) “And there are many instances of English speakers referring to themselves as *yours truly*.”

(pg. 18) “Naturally occurring examples of 1st person plural reflexives anteceded by plural imposters are frequent on the Web.”

(pg. 20) “However, even though (11) accurately represents our dialect, the Web provides numerous examples that arguably illustrate a different variant of English.”

(pg. 21) “Many other such examples are found on the Web.”

(pg. 158) “Instances of this pattern are easily found on the Web.”

(pg. 241) “We have also found one occurrence of this kind of sentence on the Web.”

David Pesetsky has pointed out a technical issue in counting hits (personal communication): “On a more technical note, if you ever care about the number of attestations — even informally to make a claim like ‘has more than 7,000 Google hits’, click through at least 30 screenfuls before believing the number of hits Google lists initially. There is a weird longstanding bug where Google will initially list a very high number of hits, but if you keep clicking through the pages, it actually turns out to be much much less. I read somewhere that the bug has something to do with how the webcrawler identifies duplicates. In any case, I've seen numbers in the thousands winnowed down to numbers like 40 once you get to page 5, for example. It's very very common. And I have seen syntax papers fall into the error of believing the initial hit number and citing it.”

To give an example illustrating the issues involved, consider again the search in (1). When I did the search on Google, Google noted “About 1,360,000 results…”. (March 6, 2023, but the number varied each time I did the search). After counting the actual hits by hand, I found only 73. Of the 73, many of them contained a result that was longer than the string I was looking for, such as “was done on my own time…”. Excluding those gave 41 hits. Four of those were duplicates (the exact same website). Excluding the duplicates gave 37 hits. I then manually looked through each of the remaining hits (opening the URL and locating the example on the website). For three of the hits, I could not find the string on the website. Excluding those gave 34 hits. The rest were of very high quality (conforming to the quality controls above).

The number of hits, and the quality of the hits, in addition to the fact that all the examples are perfectly grammatical to me, gives me confidence in the descriptive generalization: the pronoun *my* in the expression *on my own* can refer back to the implicit argument in a passive formed with the participle *done*. Further searches would broaden this generalization in various ways.

I am not suggesting that the reader go through such a painstaking counting process for their own research. For example, making sure there are no duplicates is quite painstaking work. There is really no need for such a detailed count. There is no need at all to cite specific numbers.

Citation

When I use sentences like (2) in a paper, I cite them with the URL:

(3) None of this was done on my own. To me success always takes collaboration. (https://voyageminnesota.com/interview/check-out-jess-pratts-story/)

It is also possible to cite the URL in a footnote. But one way or the other, it should be cited. That way a reader can look it up on their own, and verify the data.

Questions and Answers:

In order to further clarify the method, I pose some possible questions about the method and answer them here.

Q: Is the goal of the Internet searches to describe the English data found on the Internet?

A: No. The goal is definitely is not to give a description of the English data found on the Internet. That task is hopelessly obscure. First, there are different dialects of English represented there (e.g., varieties of American, Canadian, British South African, Indian, Australian and Ghanaian English, to name just a few). Second, the people using English are from very different ages and backgrounds (socioeconomic and cultural). Third, the English found on the internet is of all kinds of registers and styles. Fourth, there may be citations from earlier stages of English. Fifth, as noted above, a lot of the English on the internet has been produced by non-native speakers. There are just way too many dimensions to think that a description could ever be give of that data.

Q: If you find a construction attested on the internet, can you immediately conclude that it is part of the dialect of some speaker?

A: No, you cannot. There are all kinds of reasons that a construction may appear on the Internet. In fact, it is not unlikely that it was written by a non-native speaker. This is the reason for for "Comparison to Target" above. If the construction that you find is relevant to your research, you need to find native speakers that accept it.

My response here supports Jason Merchant's dictate: "Beware the fetishization of attestation!". Jason explains the phrase as follows (personal communication): "I've used it in various talks and handouts to warn people against taking attested sentences as direct input to theorizing. Particularly when a speaker might actually classify that sentence as unacceptable…The broader point is really about corpus linguistics, and why data from corpora still need to be checked with speakers, and some of it should be rejected."

Q: If during searches, you find some construction (that you find acceptable) that apparently contradicts your analysis, can you discard it because you found it on the Internet?

A: Definitely not. You must either show that it does not in fact conflict with your assumptions, or modify or completely reject your analysis. It is not licit scientifically to cherry pick the data that

you find on the Internet. Any data that you find that bears on your analysis must be treated seriously, whether or not it supports your analysis.

Q: If you search for a particular construction, and do not find it, can you use this as evidence that the construction is unacceptable?

A: No, you cannot. There are endless reasons why a particular sentence may not yield any hits in a search. To determine if a sentence is acceptable or not, you need to judge it for acceptability.

Conclusion

Even though in this blog post I have addressed syntactic research only, the same exact points carry over to semantics research and syntax/semantics interface research.

References

For further information on Internet searches in syntactic research, see:

Collins, Chris and Paul Postal. 2012. *Imposters*. MIT Press, Cambridge, MA.

Collins, Chris and Paul Postal. 2014. *Classical NEG Raising*. MIT Press, Cambridge, MA.

Collins, Chris. 2022. Principles of Argument Structure: A Merge-Based Approach. Ms., NYU. (https://ling.auf.net/lingbuzz/006409)

Collins, Chris. 2018. Using * in Google Searches for Syntactic Research. Ordinary Working Grammarian [Blog]. (https://ordinaryworkinggrammarian.blogspot.com/2018/10/using-in-google-searches_5.html)