Using linguistic corpora to study grammars in situations of variation and change

Beatrice Santorini
University of Pennsylvania
beatrice@upenn.edu

Poster Day, English Linguistics
Universität des Saarlandes
February 2, 2021

The usual way we study grammars is by eliciting grammaticality judgments.

Some caveats …

- Age of acquisition
- Dialect differences
- Interference from prescriptive grammar and other ideological pressures
- Metalinguistic talent

Even with the caveats, the advantages of judgment elicitation are obvious.

Compared to other ways of getting data, like experiments, judgments are quick, easy, cheap, …

What's not to like?

But sometimes judgments are not the right way to go.

For instance …

- Sociolinguistic variation and change

  Ordinary judgments are almost certain to be unreliable due to factors including prejudice, pure ignorance, confabulation, and so on.

- Historical variation and change

  There is no chance of getting ordinary judgments.

But we still want to do research on these topics.

What can replace judgments as our source of data?

In study of synchronic variation and change, we can turn to experiments.

But that's not going to work for studies of historical variation and change.

This is where corpora come in.

Different types of corpora

- Written historical texts
- Audio recordings of speech
- Video recordings of sign
- Additional annotation of various sorts

```
( (IP-MAT (code <DCB_se1_ag4_f_01_xmin=1405.67432>)
        (CONJ-TEMP and)
        (code <$$DCB_se1_ag4_f_01_xmax=1406.04249>)
        (code <DCB_se1_ag4_f_01_xmin=1406.15854>)
        (CP-ADV (C as)
            (IP-SUB (NP-SBJ (PRO we))
                (BED 0)
                (VAG praying)))
        (NP-SBJ (PRO we))
        (MD could)
        (VB see)
        (IP-ECM (NP (NS people))
            (VAG running)
            (code <$$DCB_se1_ag4_f_01_xmax=1408.53564>)
            (code <DCB_se1_ag4_f_01_xmin=1409.09466>)
        (ADVP (ADV outside))
        (PP (P with)
           (code <$$DCB_se1_ag4_f_01_xmax=1409.86702>)
           (code <DCB_se1_ag4_f_01_xmin=1409.96306>)
           (NP (NP (NS televisions))
              (punc ,)
              (CONJP (NP (NS sofas)))
              (CONJP (CONJ and)
                  (NP (N stuff)
                     (PP (P like)
                        (NP (D that))))))))
(punc ,)
(paren (PRO you) (VBP know))
(punc .)
(code <$$DCB_se1_ag4_f_01_xmax=1412.99646>)) (ID AAE,.57620))
```

```
( (IP-MAT (code <DCB_se1_ag4_f_01_xmin=1417.12113>)
        (NP-SBJ (PRO I))
        (VP (ADVP (ADV just))
            (DOD did@)
            (NEG @n't)
            (VB understand)
            (CP-QUE-SUB (WNP-1 (WPRO what))
                        (IP-SUB (BED was)
                                (NP-SBJ (PRO we))
                                (VP (VAG learning)
                                    (NP-ACC *T*-1)
                                    (PP (P from)
                                        (NP (D that)))))))))
(punc ,)) (ID AAE,.57622))
```

```
( (IP-MAT (code <DCB_se1_ag2_m_01_xmin=4144.74125>)
        (NP-TMP (Q Every)
                (N time)
                (CP-REL (IP-SUB (NP-SBJ (PRO I))
                                (GTP get)
                                (PP (P around)
                                    (NP (PRO you))))))
        (NP-SBJ (PRO I@))
        (BEP @'m)
        (VAG thinking)
        (PP (P about)
            (CP-QUE-SUB (WADVP (WADV how))
                        (IP-SUB (NP (PRO you))
                                (ASP done)
                                (VBD hung)
                                (NP (PRO$ my) (NS ancestors)))))
        (punc .)
        (code <$$DCB_se1_ag2_m_01_xmax=4147.93592>)) (ID AAE,.49824))
```

Some (audio-aligned) parsed corpora

Audio-aligned Parsed Corpus of Appalachian English (AAPCAppE) - https://aapcappe.commons.gc.cuny.edu/ - completed

Corpus of Regional African American Language (CORAAL) - https://oraal.uoregon.edu/coraal - corpus is growing, parsing of core subcorpora in progress

Corpus of New York City English: Audio-Aligned and Parsed (CoNYCE) - https://conyce.commons.gc.cuny.edu/ - corpus is complete, parsing in progress

Eric Haeberli and Manuela Schönenberger are compiling a corpus of spoken Swiss German from Wil. 800K words are parsed, with 200K still to be added.

A conventional way to use corpora is to use them in essentially the same way as we use consultants.

Instead of asking “Can you say this?” or “How do you say this?”, we ask “Does this phenomenon occur in the corpus?”.

- If it does, we count that as an “ok” or “grammatical”.
- If it doesn’t, we need to decide whether that means “*” or whether the absence is expected because the phenomenon is rare, in which case we’re out of luck.

But using corpora in this way still leaves questions concerning variation and change that we can’t answer.

Some thought experiments concerning phrase structure change from OV to VO

- Imagine you're studying a language that is undergoing phrase structure change from OV to VO.
- Imagine further that the grammars are very simple and the surface word order patterns stand in a one-to-one relation to the grammars that generate them.
- In other words, O-V word order unambiguously reflects the OV grammar, and same for V-O word order and the VO grammar.
- In this simple case, you can simply count up instances of O-V and V-O word order patterns over time and that will show you the replacement of the OV grammar by the VO grammar.

Yes, that's how it would work if the world were made to our order.

You might have noticed that it isn't.

Many languages with an OV grammar allow phrases to follow the verb, at least during the change in progress, leading to surface V-O orders.

So then you can't tell by looking at an instance of V-O word order whether it was generated by the VO grammar or by the OV grammar with subsequent extraposition.

Mutatis mutandis for the VO grammar and scrambling.

It doesn’t matter what we call these complicating grammatical processes or options (extraposition, exbraciation, scrambling, Nachfeld/Mittelfeld-Besetzung).

It doesn’t matter whether we adopt Kayne’s Linear Correspondence Axiom concerning head-final structures.

All that matters for present purposes is that we have reason to believe that the grammar includes alternatives that destroy the unambiguous relationship between surface word order and the grammar generating the word order.

Ok, so we have a problem. For instance, we come across a sentence in Middle English with an V-O word order, and we don’t know whether it was generated by the old OV grammar (inherited from Old English and more distantly probably from Indo-European) or the new VO grammar (which perhaps arose through contact with Old Norse).

But what if the problem were purely technical.

Maybe the problem could be solved if we had a time machine!

In that case, we could just travel back a few centuries and simply ask the person who wrote the sentence.

They know, right?

After all, they wrote the sentence!

So we remember that Mark Twain wrote a book about traveling to the court of King Arthur.

We get in touch with his heirs, and we rent his time machine for the weekend, and off we go.

Ok, we've landed in the Middle Ages, and we've finagled an introduction to Geoffrey Chaucer and we've explained about OV and VO grammars, and we've just asked him whether that sentence of his in our corpus that we're unsure about was generated using the OV grammar or the VO grammar.

He thinks for a moment and then replies - mit dem Brustton der Überzeugung - "The VO grammar".

Great - well, that answers that question.

Just before we head back to the 21st century, we do our due diligence and ask a follow-up question.

“Master Chaucer, is ’t possyble that the old grammaire and derived the V-O order using extraposicioun?”

Chaucer says, “Sure, I might have. But I assure you - by my troth - I did not so.”

All right, I hope you agree with me here that renting the time machine was a waste of money for the intended purpose.

We would be nuts to take anything at face value that Chaucer says about which grammar he used.

Assuming a change in progress, it is simply impossible to tell for any individual sentence which grammar was used to produce it.

Nevertheless, not all is lost.

It turns out to be possible to estimate the aggregate incidence of sentences produced by the OV grammar vs. the VO grammar in a corpus.

So, we can't tell with certainty for individual sentences, but we can estimate in the aggregate.

Let’s begin by assuming that the sentences in the corpus are all produced by the new VO grammar.

Under this assumption, all O-V sentences must be the result of scrambling.

(We could make the opposite assumption - OV grammar with V-O sentences as the result of extraposition - it doesn’t matter for the purposes of the argument.)

As it stands, this conclusion has no empirical content (it follows necessarily from the assumption).

Can we give it some empirical content?

Yes!

Let's find all the sentences in the corpus with the following word order patterns.

a. V-IO-DO, V-DO-IO
b. IO-V-DO
c. DO-V-IO
d. IO-DO-V, DO-IO-V

In these sentences, given our assumptions, preverbal DO's must have scrambled. We can estimate the rate of DO scrambling using the following formula:

(c+d) / (a+b+c+d)

For the sake of argument, let's say that our estimate of DO scrambling comes out as 0.12.

Let's now change our focus from ditransitive sentences to monotransitive sentences.

These will either exhibit the surface order O-V or V-O.

Let's say we have a total of 1,000 such sentences in the corpus.

If they were all produced by the VO grammar (as we are assuming), we would expect 120 (= 1,000 * 0.12) of them to exhibit O-V order (derived by scrambling), and the remaining 880 to exhibit V-O order.

Let’s say that in fact we find 600 O-V sentences and 400 V-O sentences (compared to expected 120 vs. 880).

We would conclude that our initial assumption - that all the sentences in the corpus reflect a VO grammar - is wrong. (The form of the argument is reductio ad absurdum.)

We would conclude further that at least some of the O-V sentences were produced by an OV grammar (our best estimate would be 480 = 600 observed - 120 expected O-V from VO by scrambling).

In order to obtain even better results, we could make the converse assumption (that all the sentences are generated by an OV grammar). We would then calculate an estimated rate of extraposition (a+b)/(a+b+c+d), and see whether that rate gives a better fit to the data.

In case neither assumption gives a perfect fit, we could assume that the sentences in the corpus reflect a mix of OV and VO grammars and calculate the mix of grammars that best fits the observed word order patterns. 40% OV vs. 60% VO? 10% OV vs. 90% VO?

Finally, we could use statistical tests to quantify our confidence in our conclusions.

If you're interested in this sort of reasoning being carried out on real data and not just as a thought experiment, the most beautiful and compelling case study that I know is:

Taylor, Ann. 1994. The change from SOV to SVO in Ancient Greek. Language variation and change 6:1-37.

And here's some work that applies the quantitative approach just described to historical English data and combines it with experimental studies on English and German:

Speyer, Augustin. 2010. Topicalization and stress class avoidance in the history of English. Topics in English Linguistics 69. DeGruyter Mouton.

Conclusion

Corpora are better than time machines!