Detecting Grammatical Properties in Usage Data

Anthony Kroch and Beatrice Santorini
University of Pennsylvania

A well-known limitation on the utility of corpus data for linguistic research is the absence of negative evidence, just the evidence that is readily available in the data of acceptability judgments. Of course, in the case of historical investigations, judgment data is simply unavailable. In this situation, it is tempting but dangerous to assume that non-occurring configurations are ungrammatical. A better approach, widely adopted with the growing availability of digital corpora, especially annotated ones, is to make use of the frequency information in the corpora to infer properties of the grammars underlying observed usage patterns. The most obvious patterns are diachronic developments in which a form either arises from nothing or disappears, and it has always been assumed that these cases reflect grammatical change. A more ambitious use of frequency information has been work in the spirit of Kroch's"Constant Rate Effect" (Kroch 1989). In such work, evidence is assembled to show that distinct linguistic environments sharing a common innovative grammatical feature will evolve together over time. The CRE has by now been replicated sufficiently often to be accepted as reliable. Most recently, for example, Zimmermann (2015) has carried out a large scale replication involving a dataset of more than 50K instances of the English *do*-support environment.

Less well-known than the CRE is the pattern reported in Santorini (1993), Taylor (1994) and elsewhere where we see that grammatical options (for example, "extraposition") that are not undergoing change tend to be stable in their frequency of use in corpus data. This work also reports that grammatically independent options, like the extraposition of one XP or another or both in a clause that contains two such phrases, tend to be statistically independent; that is to say, if the probability of one occurrence of extraposition is p, then the probability of two occurrences will be approximately $p \cdot p$. Largely ignored in the literature, however, is the contrapositive implication that when options are statistically linked, we have evidence of grammatical linkage.

In this paper, we present evidence from four languages for which we have parsed historical corpora: English, French, Icelandic and Yiddish (Kroch and Taylor 2000, Martineau and et al. 2009, Wallenberg et al. 2011, Santorini 2008) of statistical linkage with grammatical implications and also of the loss of such linkage over the course of time. The data on which we rely is word order inside VP, where we find that these languages undergo a shift from XV to VX in multiple stages, two of which can only be distinguished by the presence versus absence of statistical linkage between different word order options. The pattern we have found, stated within the framework of antisymmetric syntax, is that XV surface word order has sources in leftward movements of two distinct types: (1) remnant scrambling of VP with the verb itself stranded in v and (2) scrambling of an XP argument/adjunct with VP remaining *in situ*. Since, under option 2, more than one XP may scramble and since, under option 1, XPs can be stranded after the verb via a sequence of XP scrambling followed by remnant VP scrambling, an identical range of surface orders is produced by the two options. Only quantitative evidence allows us to distinguish them. Concretely, we find the following quantitative patterns in our languages:

1. From their earliest attested periods, the languages exhibit leftward movement of single XPs across the verb, as expected under the XP scrambling option.

2. At the same time, in the earlier periods, the frequency of verb-final order in clauses with multiple XPs in pre-verbal position is much higher than expected, given the frequencies of single XP movement.

3. As reported for Ancient Greek in Taylor (1994), the frequencies of leftward movement of single XPs of a given syntactic type are largely independent of the presence of other XPs in the clause.

4. After initial periods with an excess of multiple XP in pre-verbal position, the frequency of XP>V orders declines in all four languages to that predicted by the rates of single XP scrambling.

From these results, we conclude that the loss of surface OV order in our languages proceeds in three stages. In the first, which antecedes our earliest records, the remnant scrambling of VP begins to be lost, leading to an alternation between XP>V and V>XP surface orders. At this time, XP>V order in clauses with one VP-internal constituent becomes ambiguous between a VP-movement derivation and one in which single XPs scramble leftward. In the second stage, the VP-movement option disappears so that XP>V order is always derived by XP scrambling. Finally, XP scrambling itself disappears or becomes restricted to quantificational expressions.

References

Theresa Biberauer. 2003. Reconsidering the EPP and Spec-TP in Germanic. In Luisa Astruc and Marc Richards, editors, *Cambridge Occasional Papers in Linguistics (COPiL)*, number 1, pp. 100–120. Department of Linguistics, University of Cambridge.

Roland Hinterhöltzl. 2006. *Scrambling, Remnant Movement, and Restructuring in West Germanic*. Oxford University Press, Oxford and Cambridge, MA.

Anthony Kroch. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change*, 1:199–244.

Anthony Kroch and Ann Taylor. 2000. *Penn-Helsinki Parsed Corpus of Middle English*. http://www.ling.upenn.edu/hist-corpora/, second edition.

France Martineau and Anthony Kroch et al. 2009. *Corpus MCVF, Modéliser le changement: les voies du français*. University of Ottawa, first edition.

Beatrice Santorini. 1993. The rate of phrase structure change in the history of Yiddish. *Language Variation and Change*, 5:257–283.

Beatrice Santorini. 2008. Penn Yiddish Corpus. Contact author for access.

Ann Taylor. 1994. The change from SOV to SVO in Ancient Greek. *Language Variation and Change*, 6: 1–37.

Joel C. Wallenberg, Anton Karl Ingason, Einar Freyr Sigurðsson, and Eiríkur Rögnvaldsson. 2011. *Icelandic Parsed Historical Corpus (IcePaHC)*. University of Iceland. version .9.

Richard Zimmermann. November 2015. A syntactic change with lots of data: The rise of 'do'-support with possessive 'have' in american english. Manchester LEL research seminar.