The Web in Theoretical Linguistics Research: Two Case Studies using the Linguist’s Search Engine

Philip Resnik, Aaron Elkiss, Heather Taylor, and Ellen Lau
University of Maryland

Presented at the 31st Meeting of the Berkeley Linguistics Society, February 2005

In recent years a growing number of researchers have investigated the use of introspective judgments underlying work in theoretical linguistics. As Bard et al. (1996) put it, each linguistic judgment is a “small and imperfect experiment”; Schütze (1996) and Cowart (1997) discuss how instability and unreliability in such informal methods can lead to biased or even misleading results (cf. Abney, 1996; Sorace and Keller, 2005). Tools for searching naturally occurring text potentially provide an additional window on linguistic data (Christ, 1994; Corley et al., 2001; Blaheta, 2002; Kehoe and Renouf 2002; König and Lezius, 2002; Fletcher 2002; Kilgarriff 2003), but typically involve computational sophistication, restriction to a preprocessed corpus, or shallow string-oriented search, making them unattractive to “the Ordinary Working Linguist without considerable computer skills” (Manning, 2003).

We designed the Linguist’s Search Engine (LSE) to empower ordinary working linguists to search the Web for linguistic data (Resnik and Elkiss, 2003). The LSE’s architecture permits efficient search based on syntactic and lexical criteria. Using its “query by example” interface, linguists interact by saying “find sentences like this” — e.g., typing (1a) automatically produces a graphical version of (1b), which the linguist edits with a few mouse clicks to get the structure (1c), and submits to find examples like (1d). Crucially, users need not learn a query language, nor agree with (or even understand) the LSE’s automatic parse, in order to find sentences with parses similar to the exemplar.

Having described the LSE, we present two case studies of its use for mainstream linguistics research. Both cases involve linguists having no previous LSE experience, and both demonstrate its value when used in conjunction with standard linguistic methods.

First case study. Taylor (2004) developed a new analysis of comparative corollatives (2a), demonstrating that they pattern similarly to conditionals under a conventional UG account, contrary to claims that they are peripheral or odd (McCawley, 1988; Culicover and Jackendoff, 1999). LSE searches like (2b) played a central role in generalizing from the observed pre-theoretic pattern (2c) to a hypothesized structure (2d) and general theoretical account. In particular, Web data revealed a much richer range of possibilities for XP than would otherwise have been considered, confirmed via conventional informant judgments.

Second case study. Kazanina et al. (2004) investigated the influence of Principle C in on-line processing of backwards anaphora (3a). An initial reading time study exploiting the “mismatch effect” (van Gompel and Liversedge, 2003) bore out the hypothesis that predictive processing would fail to take place with the anaphor in a Principle C context; however, the stimuli confounded Principle C contexts with optional subordinate clauses (3b). The LSE played a central role in quickly identifying and validating stimuli that avoided the confound (3c), as later confirmed via a small off-line completion study.
(1) a. Exemplar for search:
   The bigger the house, the higher the price
b. Automatic parse:
   \[[x [x [x [the] [on bigger]] [or] [the] [on house]]] \]
c. User-edited tree (six mouseclicks later) specifying structure to search for:
   \[[x [x [x [the] [JJR]]] [or] [the] [JJR]]\]
d. Examples of real query results:
   The higher the rating, the lower the interest rate that must be paid to investors.
   The more you bingo, the more chances you have in the drawing.
   The more we plan and prepare, the easier the transition.
   Overall, the lower the ratio ..., the more favorably a lender views the risk of the loan.

(2) a. The louder the party is, the more the neighbors complain.
b. \[[s [s [x [The] [JJR]]] [x [the] [JJR]]]\]
c. The -er/more/less X, the -er/more/less X
d. \[[the -er/more/less XP] IP\] [[the -er/more/less XP] IP]

(3) a. *He, was fed up when the boy, visited home.
b. He was cooking dinner while John listened to the radio
c. It was clear to him/his mother that John should go.

---


