The Web in Theoretical Linguistics Research: Two Case Studies Using the Linguist's Search Engine*

PHILIP RESNIK, AARON ELKISS, ELLEN LAU and HEATHER TAYLOR
University of Maryland, College Park

0. Introduction

The whisper of "does that sound ok to you?" is a familiar sound to most linguists: we often hear it in the audience when a presenter supports a theoretical point using a judgment of grammaticality or ungrammaticality, and someone in the audience disagrees with the judgment. In recent years a growing number of researchers have investigated the use of introspective judgments underlying work in theoretical syntax and in linguistics more generally. As Bard, et al. (1996) put it, each linguistic judgment is a "small and imperfect experiment." Schütze (1996) and Cowart (1997) provide detailed discussions of instability and unreliability in such informal methods, which can lead to biased or even misleading results.

Alternatives to introspective judgments include psychological methods (Bard et al. 1996), quantitative modeling in order to approximate judgments (Lapata et al. 2001), and broader reexamination of the idealizations underlying linguistic theory (Abney 1996; Sorace and Keller 2005). Tools for searching naturally occuring 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, Evans et al. 2003), but typically these tools involve computational sophistication, restriction to a particular corpus, or shallow word-level search, making them less attractive to "the Ordinary Working Linguist without considerable computer skills" (Manning 2002).

We designed the Linguist's Search Engine (LSE) to empower ordinary working linguists to search the Web for linguistic data (http://lse.umiacs.umd.edu; Resnik and Elkiss 2005). The LSE's architecture permits efficient search using syntactic and lexical criteria, with special attention to the needs of linguists, and

* The Linguist's Search Engine project has been supported in part by the National Science Foundation, ITR grant IIS0113641, and by the Center for the Advanced Study of Language, TTO32. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation or the Center for the Advanced Study of Language.

with a clear understanding of the need to manage the tradeoff between rapid interaction and Web-scale comprehensiveness.

In Section 1, we briefly describe the Linguist's Search Engine and how it makes searches of that kind possible. Sections 2 and 3 focus on two case studies using the LSE for mainstream linguistics research, one in syntax, and the other in psycholinguistics. Both studies involved linguists with no previous LSE experience, and both demonstrate its value when used in conjunction with standard linguistic methods. Section 4 provides a discussion and conclusions.

1. The Linguist's Search Engine

As an example of the need for a tool like the LSE, consider a typical instance in which a theory makes a prediction that might or might not be valid. Resnik (1996) presents an account of implicit objects in English that hinges on a quantitative model of selectional preferences: assuming certain aspectual criteria are met, verbs are predicted to permit implicit object usages when the verb selects strongly for the semantic category of its object (e.g. *John ate*, versus **John found*). The verb *titrate* was proposed to the author as a potential counterexample – can one say *They titrated*? Without a set of informants familiar with chemistry, this is a difficult question to answer. Looking for attested usages in a large corpus such as the British National Corpus might be an alternative, facilitated by a linguistically informed interface such as VIEW (Davies 2006), but "large" is a relative term – in its 100 million words of text, the BNC contains only 27 instances of *titrate* in any inflection. Looking for attested usages on the Web might be an alternative, but standard search engines are little help: a search for transitive uses could be approximated by searching for *titrate* followed by determiners like *the* or *a*, but a search for apparently *intransitive* uses is impossible. What is needed is a way to look for data on the Web in a linguistically informed way: that is, a search engine that permits the user to say something like "Find me sentences containing a verb phrase headed by any inflection of *titrate*, such that there is no NP complement of the verb."

This example illustrates two main requirements. First, linguists need a way to specify searches that is simultaneously easy to use and linguistically sophisticated. Second, they need to be able to search on the scale of the World Wide Web, but also to examine search results quickly and modify their queries if necessary.

In order to address the first requirement, the LSE adopts a strategy we call "query by example." Figures 1 and 2 show how a user might type the sentence *John titrates the solution* into the LSE's query-by-example interface, resulting in an automatically generated parse tree (also shown as (1a)). With a few mouse clicks, the user can reduce that structure to just the parts that are of interest, shown in (1b), then specify that the NP should be *absent* rather than present, and request all morphological variations of *titrate* when used as a verb. The query in (1c) is generated automatically from the graphical structure.

Figure 1: Querying by example

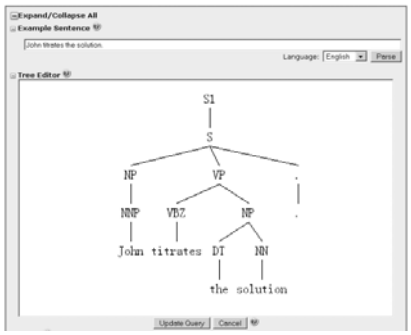
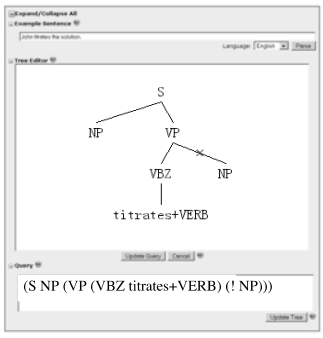


Figure 2: Resulting query



(1) a. (S1(S(NP(NNP John)) (VP(VBZ titrates) (NP(DT the)(NN solution))) (..)))
b. (S NP (VP (VBZ titrates) NP))
c. (S NP (VP (VBZ titrates+VERB) (! NP)))

Crucially, users need not learn the query language, although advanced users can edit or create queries directly if so desired. Nor do users need to agree with (or even understand) the LSE's automatic parse, in order to find sentences with parses similar to the exemplar; what is important is that the search engine will be finding sentences containing the *same* automatically generated structure, whether or not that structure is consistent with any particular linguistic theory.

The second requirement of the LSE is to manage the tradeoff between rapid response time and the ability to search the Web as a whole. By default, users search against a static collection of several million sentences sampled randomly from the Web, and this collection is often useful by itself. In order to search the entire Web, the LSE permits users to define their own custom collections, taking advantage of standard commercial Web search engines. To search for instances of *titrate* using the LSE's "Build Custom Collection" functionality, the user can specify that the LSE should do the following:

- Use standard Web search to find pages with any morphological form of *titrate*
- Extract only sentences containing that verb
- Automatically parse and index those sentences
- Augment the collection by iterating this process with different specifications

Doing the Web search and extracting, parsing, and indexing the sentences can take some time, but the LSE permits the user to begin searching his or her custom collection as soon as *any* sentences have been added into it. Typically dozens to hundreds of sentences are available within a few minutes, and a typical custom collection, containing thousands or tens of thousands of sentences, is completed within a few hours. A custom collection for *titrate* contains 3831 sentences, and

the query in Figure 2 and (1c) returns 145 matches, including numerous uses that attest to its behavior as an implicit object verb, some of which are illustrated below.

(2) a. You titrate with sulfuric acid using gloves and safety glasses
b. We titrated till we reached the final endpoint of pH 4.5.
c. Conversely, if we titrate in the opposite direction…
d. I no longer titrate, but that is because…

As a cautionary note, many of the sentences that match the specified syntactic structure do *not* represent implicit object uses.

(3) a. Note that the detected response titrates linearly
b. The catheter is inserted and the optimal dose titrated.
c. Although sulfite usually titrates as thiosulfate…

This demonstrates the danger in simply counting the number of matches rather than examining the data. When data are returned by LSE searches, a responsible linguist must consider the context, whether the sentence originated with a native speaker, and so forth – just as when querying informants for linguistic judgments. In contrast to informant judgments, though, LSE results come with pointers to the Web pages where the sentences occurred, so that other linguists can evaluate the quality of the data for themselves.[1]

2. Syntax: Using the LSE to Investigate Comparative Correlatives

Taylor (2004) describes an investigation of comparative correlatives (also known as "comparative conditionals" or "conditional comparatives," henceforth CCs), illustrated by the example in (4). CCs consist of two clauses, which are labeled here, with no specific theoretical bias, as clause_1 and clause_2.

(4) The more pizza Romeo eats, the fatter he gets
clause_1 clause_2

CCs have been highlighted recently in the debate over whether Universal Grammar (UG) can reasonably account for the acquisition of all naturally occurring linguistic data (McCawley 1988, Culicover and Jackendoff 1999, Culicover 1999, Borsley 2003, den Dikken 2004, Goldberg and Jackendoff 2004). In particular, Culicover and Jackendoff (1999) argue that syntactically, CC constructions are *sui generis*, behaving at odds with accepted views of general UG constraints on syntax.

Taylor used the LSE in order to explore a more comprehensive range of data on CCs than previously discussed in the literature, with interesting results. The

[1] See http://umiacs.umd.edu/~resnik/bls2005_data.html for an online appendix to this paper.

first result concerns the fact that CCs can occur with optional deletion of a main copular verb in each clause (see (5)).

(5) The better an advisor ~~is~~, the more successful the student ~~is~~

McCawley (1988) made the generalization – accepted without challenge in the subsequent literature – that copula deletion of this kind is only licit when the subject of the clause is generic, rather than specific, as evidenced by the contrast between (5) and his example (12a), reproduced here as (6):

(6) The more obnoxious Fred is/*∅, the less attention you should pay to him

Since the distinction between generic and specific arguments is a semantic one, McCawley's generalization provides support for the idea that at least some constraints on CCs must be accounted for outside their syntax.

A search for naturally occurring CCs, however, showed that there is more to the story. Using the "query by example" process, a search of the static LSE Web collection (3.5 million sentences) yielded an unexpected result. While it was true that, in instances of copula deletion, CCs commonly occurred with generics in their subject, it was more striking that *all* instances of copula deletion included deletion of a main copular verb *in both clauses*. This observation, based on the LSE search results, suggested that the unacceptability of copula deletion in McCawley's original example (6) arises from an inability to delete or retain the copula in parallel.

Analysis of the LSE data led to consideration of an additional factor: it appears that the acceptability of copula deletion with a specific subject improves as the phonological weight of the subject increases, as illustrated in (7).

(7) a. *The more obnoxious Fred, the less attention you should pay to him
b. ?The more obnoxious a child, the less attention you should pay to him
c. The more obnoxious Fred's younger brother, the less attention you should pay to him.

On the basis of the LSE-driven observations and her own observations, Taylor obtained confirming judgments regarding parallelism from informants:[2]

(8) a. ?? The longer the day's activities last, the sleepier the campers ~~are~~
b. ?The longer the day's activities are, the sleepier the campers ~~are~~
(9) a. ?? The more tiring the day's activities ~~are~~, the more food the campers eat
b. ?The more tiring the day's activities ~~are~~, the sleepier the campers are

[2] ? ≈ ok, but odd, ?? ≈ strange, but could be uttered/understood. Informants were asked to judge (potential) contrasts between sentence pairs, rather than judging sentences in isolation.

(10) a. ?The longer the day's activities are, the sleepier the campers ~~are~~
b. ?The more tiring the day's activities ~~are~~, the sleepier the campers are
c. √The more tiring the day's activities ~~are~~, the sleepier the campers ~~are~~

The data gathered with the help of the LSE demonstrate that McCawley's generalization cannot simply be taken at face value. Whatever the underlying explanation for their role, syntactic parallelism and phonological weight clearly play a part in informants' judgments about the acceptability of copula deletion, and those factors might in fact explain McCawley's observation in (6) without making recourse to the generic/specific distinction at all.

Taylor's second result using the LSE goes back to McCawley's (1988) original term for CCs, "comparative conditional." He noted that the interpretation of CCs seems to be similar to that of standard conditionals, as in the relationship between (4) and (11).

(11) If Romeo eats more pizza, then he gets fatter

Culicover and Jackendoff (1999) chose to rename the expressions "comparative correlatives", noting that the expressions' interpretations are closer to that of "as" phrases, like (12).

(12) As Romeo eats more pizza, he gets fatter.

The question of whether CCs are akin to correlatives or conditionals cross-linguistically is not without consequence: the syntactic analysis of correlatives and conditionals, and whether UG can handle these data, may bear on whether UG can explain their acquisition. Iatridou (1991) theorizes that the IF-clause of a conditional can be base-generated in a sentence-final position low in the structure, and A'-move to a sentence-initial position above the main clause. Taylor (2004) observes that clause$_1$ of a CC behaves syntactically like an IF-clause with respect to not hosting tag questions, failing to host subjective mood, licensing NPIs in the absence of a NEG item, permitting extraction, and variable binding. Putting these ideas together, Taylor hypothesizes that clause$_1$ of a CC is base generated low in the structure, and obligatorily A'-moves to a higher, sentence-initial position.

In Taylor's exploration of that hypothesis, the LSE again made it possible to find relevant data. In discussions of CC phenomena, two colleagues reported that in their dialect of English, clause$_2$ of a CC could begin with an overt instance of *then*, as in (13).

(13) The more pizza Romeo eats, then the fatter he gets

Now, seeing clause-initial *then* show up overtly in clause$_2$ would certainly lend strong support for the hypothesis that CCs are closely akin to conditionals, providing additional motivation to pursue an analysis of CCs similar to that of

conditionals. But what would constitute sufficient evidence? Linguists' judgments can be biased, especially when arrived at in the midst of a theoretical discussion; those judgments would be much more valuable if corroborated by linguistically naïve or naturally occurring sources.

The LSE made it possible for us to quickly find such corroborating data, as illustrated in (14):

(14) a. The darker the coffee bean, then the less caffeine.
b. The more playing time in the past then the less regression to the mean needed
c. The more that you can arrange for everyday life to happen almost automatically then the less you have to concentrate on what is happening

Overt *then* may not be a particularly frequent phenomenon, and it may be specific to certain dialects. But the Web data help in establishing that overt *then* can, in fact, occur in CCs, supporting McCawley's (1988) characterization of CCs as a type of conditional, as well as Taylor's continuing treatment of CCs as such.

A full treatment of comparative conditionals, and the more general implications for UG, are still a matter for further research. But whatever one's position on this construction and UG in general, experience with the LSE demonstrates that linguistically informed Web search can complement the introspective methods of syntactic theory, *without* requiring the syntactician to expend significant effort learning and applying the machinery of alternative data-gathering methods.

3. Psycholinguistics: Using the LSE to Investigate Backward Anaphora

Lau and colleagues have recently conducted a series of psycholinguistic experiments that examine the question of how linguistic principles constrain the processes involved in online processing (Kazanina, Lau, Liberman, Phillips, and Yoshida 2004; 2005; in prep.). Specifically, they are interested in the influence of Principle C, postulated by Chomsky (1981) to govern coreference interpretation mainly in situations where a pronoun precedes a full nominal expression – configurations known as *backwards anaphora*. The rule says that the pronoun and the nominal cannot corefer when the pronoun c-commands the nominal, which captures contrasts such as (15).

(15) a. *He_i promised that John_i would go.
b. His_i mother promised that John_i would go.

Previous reading time studies have shown that *predictive processing* takes place in backwards anaphora constructions like (16a), below: the parser anticipates finding the referent for *he* in the subject position of the matrix clause (underlined), and it is therefore "surprised" when it encounters a gender-mismatching referent there as in (16b) (van Gompel and Liversedge 2003).

(16) a. While he was cooking dinner, John listened to the radio.
b. While he was cooking dinner, Mary listened to the radio.

The parser's surprise shows up in longer reading times at the gender-mismatching region, so this has become known as the "gender mismatch effect" (GME).

Lau and colleagues reasoned that if Principle C, a rule of the grammar, actively constrained the earliest parsing processes, then a GME would *not* be seen in constructions in which Principle C would rule out coreference to the target position, as in the matrix-subordinate construction, such as (17a). This null effect would contrast with the GME for the exactly matching subordinate-matrix combination in (17b). Alternatively, if Principle C were *not* immediately available to constrain the predictive parsing process, then one would predict that a GME would be seen for both constructions.

(17) a. He was cooking dinner while John/Mary listened to the radio.
b. While he was cooking dinner John/Mary listened to the radio.

Their first experiment examined such subordinate-matrix/matrix-subordinate pairs using reading time measures, and the results bore out the former prediction, i.e. that Principle C was immediately available to block the GME. However, the authors soon realized there was a confound in their experimental design: in the Principle C conditions, the subordinate clause following the matrix was optional, so the target position wasn't obligatorily predicted at the pronoun position. On the other hand, in the non-Principle C conditions, the matrix clause following the subordinate was obligatory, so the target position was guaranteed. In other words, the results they observed could have followed from the fact that predicted coreference would only occur when a suitable syntactic position for the referent was guaranteed independently by other syntactic requirements.

Lau and colleagues needed a construction that could be varied minimally to manipulate its Principle C properties and which did not have this confound: *all* conditions should syntactically require the target position. One possibility they began to consider were expletive constructions, such as *It was clear to him that…* These constructions appear to obey Principle C, in that coreference with the complement clause's subject seems to be impossible; see (18a). The sentence also can be minimally altered to provide a non-Principle C pair, as in (18b).

(18) a. It was clear to him$_i$ that John$_{*i/j}$ should go.
b. It was clear to his$_i$ mother that John$_i$ should go.

However, remember that the crucial issue for removing the confound is that all conditions should syntactically require the target position (containing the name). Intuitively, without any discourse context these sentences do seem to require a complement clause at the appropriate point – *It was clear to him/his mother…*

seems to call for a continuation – but the requirement isn't absolute, since *It was clear to him* can also be a perfectly good sentence in the right context.

Overall, then, this construction seemed promising for solving the confound, but since the prediction of the complement clause subject independent of coreference was critical, the authors needed a way to verify their intuition that the complement clause was indeed obligatory in the context of the experiment. Sifting through a large corpus would have been one solution to this problem, but this was daunting to tackle for designing a reading-time study.[3] The LSE provided Lau and colleagues with a quick and straightforward way to check their concerns about their materials.

The expletive construction which the authors were concerned with were of the form "It [aux verb] [adjective] to [NP]…" Obtaining a representative sample of constructions of this form using a standard Web search would be difficult because of the variability in the lexical items that could be inserted in these slots (e.g. *it was clear to, it is obvious to, it seemed necessary to*, etc.). Using the LSE, the authors were able to search exactly for sentences of exactly this form. The results of the search allowed them to ascertain that in virtually all cases of this construction, as attested in naturally occurring Web data, the *it* was indeed the expletive *it*; moreover, in these cases a complement clause does always follow the initial phrase. This was the evidence needed to proceed confidently with this construction in designing their follow-up experiment.

Lau and colleagues later double-checked the same point experimentally by running a small off-line completion study and showed that in the majority of cases participants do finish the sentence with a complement clause, supporting the conclusions they drew from the results of the LSE search. The mutually reinforcing results obtained by these two methods highlight the unique advantages of using the LSE in designing experimental materials: where pre-tests and norming studies require a significant investment of time and energy, an LSE search and analysis of the results can be done in an afternoon or less. Thus, doing an LSE search *before* running norms or pre-tests of materials can indicate whether the materials are promising enough to merit going the next step to norming, and can also help *refine* the materials before they are tested on participants.

Ultimately, proceeding with the expletive construction follow-up on the basis of these results, using materials that avoided the original confound, the authors showed that the GME still did not show up in the Principle C conditions, in contrast with the unconstrained conditions. Combined with the results of a third experiment (which used still another construction to avoid the confound), the authors have been able to present a strong case that the early predictive parsing processes involved in the computation of coreference are constrained by a rule governing coreference in the grammar.

[3] The authors were also facing a serious time constraint: their abstract had been submitted and accepted before they discovered the critical confound in their first experiment, and they needed to run the follow-up before the conference!

4. Conclusions

Like any other scientific enterprise, linguistics makes progress via a cycle of theoretical and empirical development. Theoretical developments lead researchers to the data, to confirm or disconfirm predictions, and data leads researchers back to their theories, to provide an account for the patterns that have been observed. There is always a risk that at some point in this process too broad or too narrow a view will be taken. Theory-driven thinking may focus so narrowly that it misses relevant data. Data analysis may cast such a wide net that relevant generalizations are lost in the noise.

Linguistically sophisticated search on the Web can help reduce these risks. By looking at naturally occurring instances, rather than just introspective judgments on constructed examples, a theorist can uncover behavior that pushes the boundaries of the current theory, or calls into question widely held assumptions. The use of the LSE in studying comparative correlatives illustrates this point: looking at dozens of naturally occurring examples, it was strikingly obvious that parallelism has an important role to play in any discussion of McCawley's theoretical point about copula deletion. How likely is it that parallelism would have emerged as a factor to be considered, given that the theory was not motivating anyone to systematically obtain examples that varied on that dimension? And how easy would it be for a typical syntactician to sample enough informants to confirm that overt *then* is a real phenomenon that any theoretical treatment of CCs needs to account for?

By the same token, large-scale data analysis benefits significantly from linguistically sophisticated search. Standard search engines are based on words or fixed phrases, so searching for a pattern like "It [aux verb] [adjective] to [NP]" is difficult, if even possible. The LSE made it possible to very quickly explore the general pattern of behavior for sentences containing this structure, confirming experimenter intuitions that they have a strong tendency for the *it* to be expletive, and for the NP to be followed by a sentential complement.

Nothing about our use of Web data runs counter to methodological standards in mainstream linguistics – naturally occurring sentences, like human judgments, require careful consideration of the context and the linguistic background of the speaker/writer/hearer. Syntactic theories can be informed by Web data and also take advantage of introspective judgments. Psycholinguistics experiments can use Web search to quickly validate experimenter intuitions, and also confirm them at greater length via completion studies and the like. What has been needed, we would argue, are tools for Web search that are sufficiently sophisticated, but still easy enough for any linguist to try – and a demonstration that, searched properly, the Web has something to offer. Unlocked by tools for searching naturally occurring text, the Web stands open as a vast, diverse, and dynamic repository of linguistic behavior, ready for exploration.

References

Abney, Steven. 1996. Statistical methods and linguistics. In Judith Klavans and Philip Resnik (eds.), *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Cambridge, MA: MIT Press.

Bard, E.G., D. Robertson, and A. Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72(1): 32-68.

Borsley, Robert. 2003. On the Polish periphery: Comparative correlatives in Polish. In P. Banski and A. Przepiórkowski (eds.), GLIP-5: Proc. 5th Generative Linguistics in Poland Conference, Polish Academy of Science, Warsaw.

Blaheta, Don. 2002. Documentation for tgrep, tsed, and wsjsed, http://www.cs.brown.edu/people/dpb/tsed/manual.pdf.

Bod, R., J. Hay, and S. Jannedy (eds.). 2002. *Probabilistic Linguistics*. MIT Press.

Chomsky, N. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Christ, Oli. 1994. A modular and flexible architecture for an integrated corpus query system. *COMPLEX'94*, Budapest.

Corley, Steffan, Martin Corley, Frank Keller, Matthew W. Crocker, and Shari Trewin. 2001. Finding Syntactic Structure in Unparsed Corpora: The Gsearch Corpus Query System. *Computers and the Humanities* 35(2):81-94.

Cowart, Wayne. 1997. *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. Thousand Oaks, CA: Sage Publications.

Culicover, Peter. 1999. *Syntactic Nuts: Hard Cases, Syntactic Theory, and Language Acquisition*. Oxford: Oxford University Press.

Culicover, Peter and Ray Jackendoff. 1999. The view from the periphery: the English comparative correlative. *Linguistic Inquiry* 30:543-71.

Davies, Mark. 2006. Variation in English Words and Phrases (website). http://view.byu.edu.

den Dikken, Marcel. 2004. Comparative correlatives comparatively. Ms., The Graduate Center of The City University of New York.

Fletcher, William. 2002. Making the Web more useful as a source for linguistic corpora. *North American Symposium on Corpus Linguistics.*

Goldberg, Adele and Ray Jackendoff. 2004. The English resultative as a family of constructions. *Language* 80:532-568.

Iatridou, Sabine. 1991. Topics on Conditionals. Ph.D. diss., MIT.

Jurafsky, Daniel. 2002. Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In Bod, et al.

Kazanina, N., E. Lau, M. Liberman, C. Phillips, and M. Yoshida. 2004. Active dependency formation in the processing of backwards anaphora. Paper presented at the 17th Annual Meeting of the CUNY Conference on Human Sentence Processing, College Park, MD, March.

Kazanina, N., E. Lau, M. Liberman, C. Phillips, and M. Yoshida. 2005. Constraints on coreference in the online processing of backwards anaphora. Poster presented at the 18th Annual Meeting of the CUNY Conference on Human Sentence Processing, Tucson, AZ, April.

Kazanina, N., E. Lau, M. Liberman, C. Phillips, and M. Yoshida. In prep. Coreference constraints in the online processing of backwards anaphora.

Kehoe, Andrew and Antoinette Renouf. 2002. WebCorp: Applying the Web to linguistics and linguistics to the Web. *Proc. WWW2002*, Honolulu, Hawaii, 7-11 May, http://www2002.org/CDROM/poster/67/.

Kilgarriff, Adam and Gregory Grefenstette. 2003. Introduction to the special issue on the Web as corpus. *Computational Linguistics* 29(3):333-348.

Kilgarriff, Adam, Roger Evans, Rob Koeling, and David Tugwell. 2003. WASPBENCH: a lexicographer's workbench incorporating state-of-the-art word sense disambiguation. *Proc. EACL 2003*, 211-214.

König, Esther and Wolfgang Lezius. 2002. A description language for syntactically annotated corpora. *Proc. COLING*, 1056-1060, Saarbrücken.

Lapata, Maria, Frank Keller, and Sabine Schulte im Walde. 2001. Verb frame frequency as a predictor of verb bias. *Journal of Psycholinguistic Research* 30(4):419-435.

Manning, Christopher. 2002. Probabilistic syntax. In Bod et al.

McCawley, James. 1988. The comparative conditional constructions in English, German and Chinese, *Proc. 14th BLS*, 176-187.

Resnik, Philip. 1996. Selectional constraints: an information-theoretic model and its computational realization, *Cognition* 61:127-159.

Resnik, Philip and Aaron Elkiss. 2005. The Linguist's Search Engine: an overview. Demonstration papers, *Proc. ACL*, Ann Arbor, Michigan, June.

Schütze, C. 1996. *The Empirical Base of Linguistics*. University of Chicago Press.

Sorace, Antonella and Frank Keller. 2005. Gradience in linguistic data. *Lingua* 115(11):1497-1524.

Taylor, Heather Lee. 2004. Interclausal (co)dependency: the case of the comparative correlative. Paper presented at Michigan Linguistics Society, University of Michigan – Flint.

van Gompel, R. and S. Liversedge. 2003. The influence of morphological information on cataphoric pronoun assignment. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29:128-139.

Philip Resnik,[1,2] Aaron Elkiss,[2] Ellen Lau,[1] Heather Lee Taylor[1]
[1]Department of Linguistics and [2]Institute for Advanced Computer Studies
University of Maryland, College Park, MD 20742

resnik@umd.edu
aelkiss@umiacs.umd.edu
ellenlau@umd.edu
hltaylor@wam.umd.edu