LAMP-TR-108
CS-TR-4541
UMIACS-TR-2003-109

NOVEMBER 2003

THE LINGUIST'S SEARCH ENGINE: GETTING STARTED GUIDE

Philip Resnik and Aaron Elkiss

Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742-3275
resnik@umd.edu

Abstract

The World Wide Web can be viewed as a naturally occurring resource that embodies the rich and dynamic nature of language, a data repository of unparalleled size and diversity. However, current Web search methods are oriented more toward shallow information retrieval techniques than toward the more sophisticated needs of linguists. Using the Web in linguistic research is not easy.

It will, however, be getting easier. This report introduces the Linguist's Search Engine, a new linguist-friendly tool that makes it possible to retrieve naturally occurring sentences from the World Wide Web on the basis of lexical content and syntactic structure. Its aim is to help linguists of all stripes in conducting more thoroughly empirical exploration of evidence, with particular attention to variability and the role of context.

Keywords: Search engines, linguistics, parsing, corpora.

This research sponsored by the National Science Foundation under ITR IIS0113641.

The Linguist’s Search Engine: Getting Started Guide

Philip Resnik[1,2] and Aaron Elkiss[2]

[1]Department of Linguistics and
[2]Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742
resnik@umd.edu

Introduction

A highly influential (some would say dominant) tradition in modern linguistics is built on the use of linguists' introspective judgments on sentences they have created. The judgment as grammatical or ungrammatical, the presentation of a minimal pair, whether or not a particular structure is felicitous given an intended interpretation – these are very often the working materials of the linguist, the data that help to confirm or disconfirm hypotheses and lead to the acceptance, refinement, or rejection of theories.

Although naturally occurring sentences are currently accorded less emphasis by many linguists, the use of text corpora has a tradition in the greater linguistic enterprise (e.g., Oostdijk and de Hann, 1994). And with the emergence of the World Wide Web, we have before us a naturally occurring resource that embodies the rich and dynamic nature of language, a data repository of unparalleled size and diversity. Unfortunately, current Web search methods are oriented more toward shallow information retrieval techniques than toward the more sophisticated needs of linguists. Using the Web in linguistic research is not easy.

The tool introduced in this getting-started guide is designed to make it easier. The Linguist's Search Engine (LSE) is a new linguist-friendly facility that makes it possible to retrieve naturally occurring sentences from the World Wide Web on the basis of lexical content and syntactic structure. With the Linguist’s Search Engine, it will be easier to take advantage of a huge body of naturally occurring evidence – in effect, treating the Web as a searchable linguistically annotated corpus.

Why should this matter? As Sapir (1921) points out, “All grammars leak.” Abney (1996) elaborates: “[A]ttempting to eliminate unwanted readings . . . Is like squeezing a balloon: every dispreference that is turned into an absolute constraint to eliminate undesired structures has the unfortunate side effect of eliminating the desired structure for some other sentence.” Moreover, Chomsky (1972) remarks that "crucial evidence comes from marginal constructions; for the tests of analyses often come from pushing the syntax to its limits, seeing how constructions fare at the margins of acceptability." It is not surprising, therefore, that judgments on crucial evidence may differ among individuals; as linguists we have all shared the experience of the student in the syntax talk who hears the speaker declare a crucial example ungrammatical, and whispers to his friend, “*Does that sound ok to you?*” The fact is, language is variable (again, Sapir, 1921) – yet in the effort to make the study of language manageable, a dominant methodological choice has been to place variability and context outside the scope of investigation.
.
While there are certainly arguments to made for focusing theory development on accounting for observed generalizations, rather than trying to account for individual sentences (perforce including exceptions to generalizations) as data, an alternative to narrowing the scope of investigation is to make it easier to investigate a wider scope in interesting ways. A central goal of our work, therefore, is to help theory development to be informed by a more thoroughly empirical exploration of real-world observable evidence, an approach that explicitly acknowledges and explores the roles of variability and context, using naturally

occurring examples *in concert with* constructed data and introspective judgments.[1] In short, to make it easier for more linguists to do the things that some linguists already do with corpora.

Now, as noted above, using corpora in linguistics is not new, and certainly there are quite a few resources available to the determinedly corpus-minded linguist (and corpus-minded linguists using them). These include large data gathering and dissemination efforts (such as the British and American National Corpora, the Linguistic Data Consortium's Gigaword corpora, CHILDES, and many others), important and highly productive efforts to annotate naturally occurring language in linguistically relevant ways (from the Brown Corpus through the Penn Treebank and more recent annotation efforts such as PropBank and FrameNet), and tools designed to permit searches on linguistic criteria (ranging from concordancing tools such as Wordsmith, Scott 1999, to tree-based searches such as *tgrep*, and beyond to grammatical search facilities such as Gsearch, Corley et al. 2001). When it comes to exploiting linguistically rich annotations in large corpora for linguistic research, however, Manning (2003) describes the situation aptly, commenting, "it remains fair to say that these tools have not yet made the transition to the Ordinary Working Linguist without considerable computer skills."

Getting Started with the LSE

The LSE is designed to be a tool for the Ordinary Working Linguist without considerable computer skills. As such, it was designed with the following criteria in mind:[2]

- **Must** minimize learning/ramp-up time
- **Must** have a linguist-friendly "look and feel"
- **Must** permit real-time interaction
- **Must** permit large-scale searches
- **Must** allow search using linguistic criteria
- **Must** be reliable
- **Must** evolve with real use

The design and implementation of the LSE, guided by these desiderata, is a subject for another document. The subject of *this* document is the first criterion. Since the LSE is a tool designed for hands-on exploration, we introduce it not by providing a detailed reference manual, but by providing a walk-through of some hands-on exploration. This is organized as a series of steps for the user to try out himself or herself – what to type, or click, or open, or close, accompanied by screen shots showing and explaining what will happen as a result.

Two words of caution. First, the LSE is a work in progress, and as such, parts of it are likely to evolve rapidly – indeed, feedback from real users trying it out should play a critical role in its further development. This means that before too long, the screen shots or directions in this guide may be out of date. If the interface is well enough designed, a user starting with this guide should still be able to explore the LSE's

[1] One can go further, to a more thoroughly probabilistic view of grammar, as suggested by Abney (1996), Manning (2003), and others. I am sympathetic to that viewpoint, and I like the way Chris Manning (2003) puts it: "To go out on a limb for a moment, let me state my view: generative grammar has produced many explanatory hypotheses of considerable depth, but is increasingly failing because its hypotheses are disconnected from verifiable linguistic data. . . I would join Weinreich, Labov, and Herzog (1968, 99) in hoping that 'a model of language which accommodates the facts of variable usage . . . leads to more adequate descriptions of linguistic competence.'" That said, I would emphasize that the LSE's main mission – to permit richer empirical investigation of naturally occurring language data – is at least compatible with linguists of all (well, most) stripes.

[2] Also worthy of note: The Robustness Principle ("Be conservative in what you do, be liberal in what you accept from others," Jon Postel, RFC 793) and The Principle of Least Astonishment ("A program should always respond in the way that is least likely to astonish the user"; one Web source attributes this to Grady Booch. 1987. Software Engineering with Ada. 2nd Ed. Benjamin Cummings, Menlo Park, CA, p. 59).

various features, even if the screen details or the exact operations have changed somewhat. But the reader should be aware of the potential discrepancies.

Second, no tool can substitute for a researcher's judgment. The LSE will, one hopes, make it easier to work with large quantities of naturally occurring data in ways that some linguists will care about. But one must be aware of all the customary cautions that come to mind when working with naturally occurring data, or with any search engine, for that matter. Questions that must be asked include things like: Is the source of this example a native speaker of English? Am I looking at written language or transcribed speech? Are the data I'm looking at providing an adequate (or adequately balanced, if that matters) sample of the language with respect to the phenomena I'm investigating? Is any particular "hit" in a search *really* an example of the phenomenon I'm looking for, or might it be a false positive?

Rather than ending with caution, though, let me end this introduction with encouragement. The LSE is a *Field of Dreams* endeavor, built on faith that "if you build it, they will come." We've built it, or at least a first version of it. Will it turn out to be a useful tool for studying language? That's a question for the readers of this document: the community of users who will, we hope, find ways to employ the LSE with insight and creativity.

Acknowledgments

It's traditional to put acknowledgments at the conclusion of a document, but it is to be hoped that momentarily the reader will be having too much fun with the LSE to pay attention to details placed at the end.

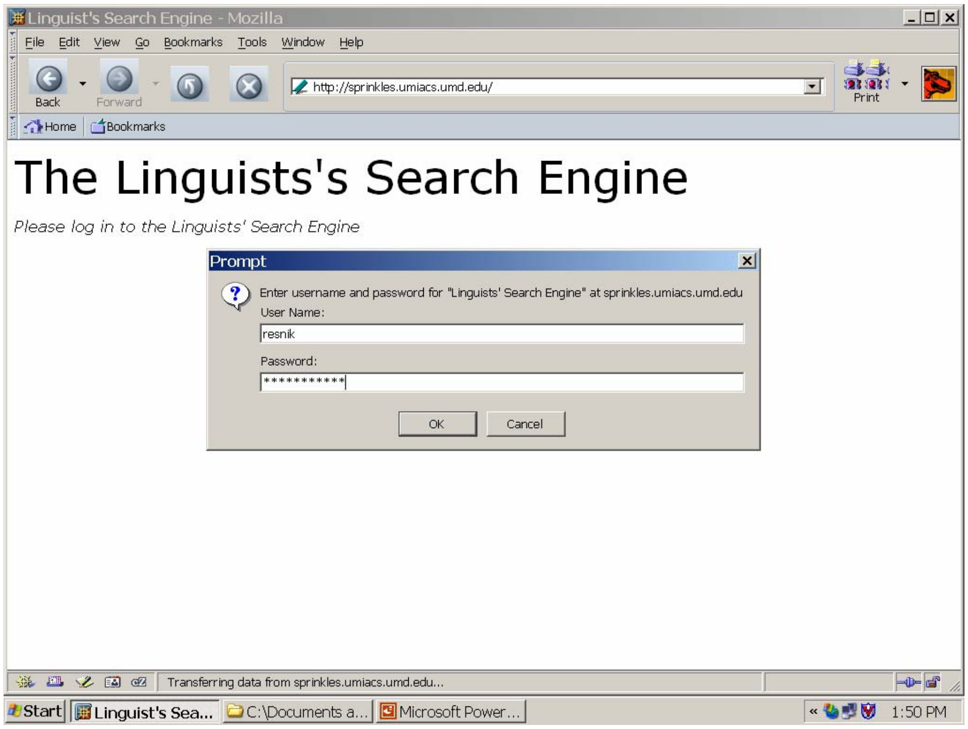
The LSE is part of a collaboration between the author and Christiane Fellbaum of Princeton University on using the Web as a source of empirical data for linguistic research, sponsored by NSF ITR IIS0113641; this collaboration also includes Mari Broman Olsen of Microsoft.

The primary implementor for the LSE is Aaron Elkiss, with contributions by Jesse Metcalf-Burton, Mohammed "Rafi" Khan, Saurabh Khandelwal, and G. Craig Murray. Critical tools underlying the LSE, without which this work would be unimaginable, include Adwait Ratnaparhki's MXTERMINATOR and MXPOST, Eugene Charniak's stochastic parser, Dekang Lin's Minipar parser (searches not currently available), Douglas Rohde's *tgrep2*, and a host of publicly available tools for construction of Web applications.

The author of this guide appreciates the early efforts and comments of the students in his spring 2003 lexical semantics seminar, which provided early feedback on a rather more preliminary version of the LSE. I am also grateful for the inspiration and lucid argumentation of empirically minded linguists Steve Abney and Chris Manning, for stimulating discussions with Bob Frank, Mark Johnson, and Paul Smolensky, and especially for the staggeringly important work of George Miller, Mitch Marcus, and Brewster Kahle (and their many collaborators) in producing WordNet , the Penn Treebank, and the Internet Archive.

I'm sure these acknowledgments are incomplete; apologies to anyone I've missed. Ditto for relevant bibliographic citations… all feedback is welcome.

First steps: Logging in and Query By Example



http://sprinkles.umiacs.umd.edu/

The Linguists's Search Engine

Please log in to the Linguists' Search Engine

Prompt

Enter username and password for "Linguists' Search Engine" at sprinkles.umiacs.umd.edu

User Name: resnik

Password: ***********

Transferring data from sprinkles.umiacs.umd.edu...

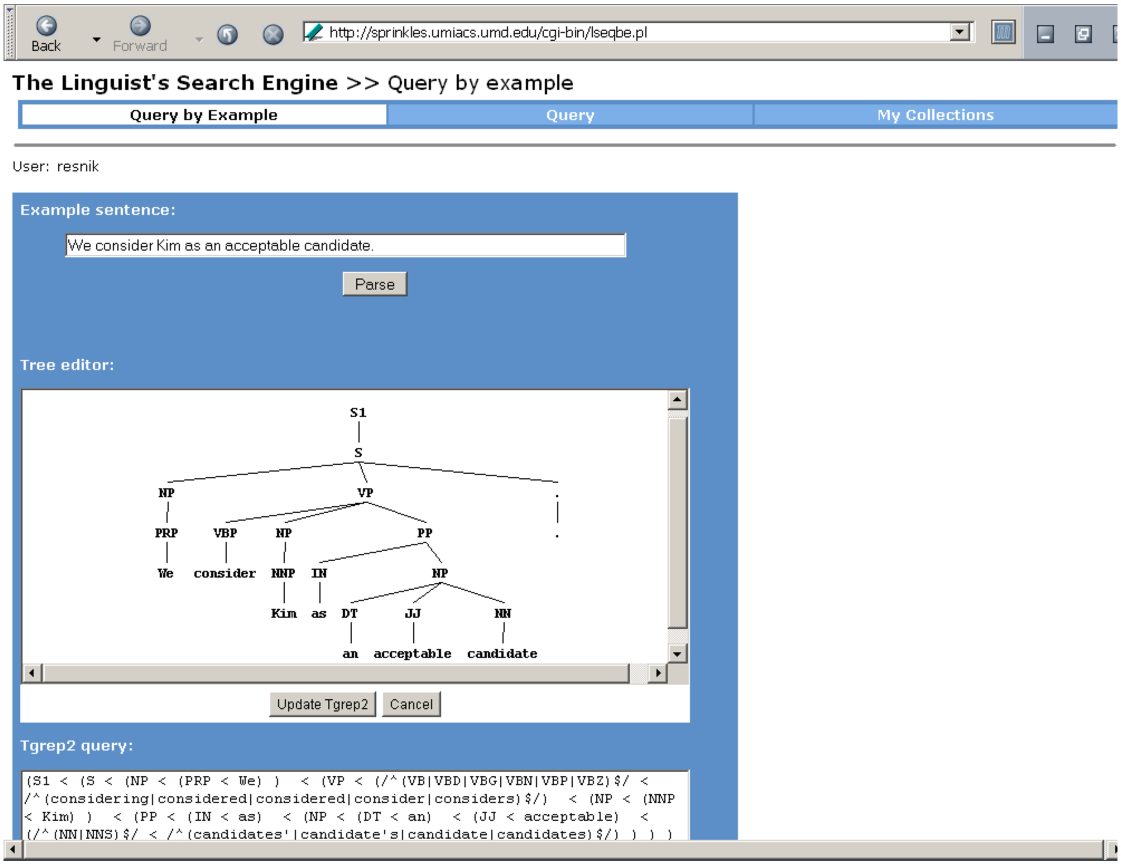
(For the impatient reader: ***focus on the instructions in bold face type.****)*

You access the LSE via your Web browser. Although a number of browsers should work, at the moment Internet Explorer (6 and higher) and Mozilla are most likely to work well. **At the entry point to the LSE, you will be asked for a login and password.** These will either have been provided to you in advance, along with the Web URL to go to, or you will soon be able to create them using a registration form. **Enter your login and password information in your browser in the usual way.**

The first example we will work with is from the discussion of Pollard and Sag (1994) in Manning (2003). The following introspective judgments are given for complements of the verb *consider*, illustrating the claim that it cannot take *as* complements.

1(a) We consider Kim to be an acceptable candidate
(b) We consider Kim an acceptable candidate
(c) We consider Kim quite acceptable
(d) We consider Kim among the most acceptable candidates
(e) *We consider Kim as an acceptable candidate
(f) *We consider Kim as quite acceptable
(g) *We consider Kim as among the most acceptable candidates
(h) *We consider Kim as being among the most acceptable candidates

Do naturally occurring data support Pollard and Sag's judgment that 1(e) cannot be used to mean the same thing as 1(a)?



The Linguist's Search Engine >> Query by example

User: resnik

Example sentence:

We consider Kim as an acceptable candidate.

Tree editor:

Tgrep2 query:

```
(S1 < (S < (NP < (PRP < We) )  < (VP < (/^(VB|VBD|VBG|VBN|VBP|VBZ)$/ <
/^(considering|considered|considered|consider|considers)$/)  < (NP < (NNP
< Kim) )  < (PP < (IN < as)  < (NP < (DT < an)  < (JJ < acceptable)  <
(/^(NN|NNS)$/ < /^(candidates'|candidate's|candidate|candidates)$/) ) ) )
```

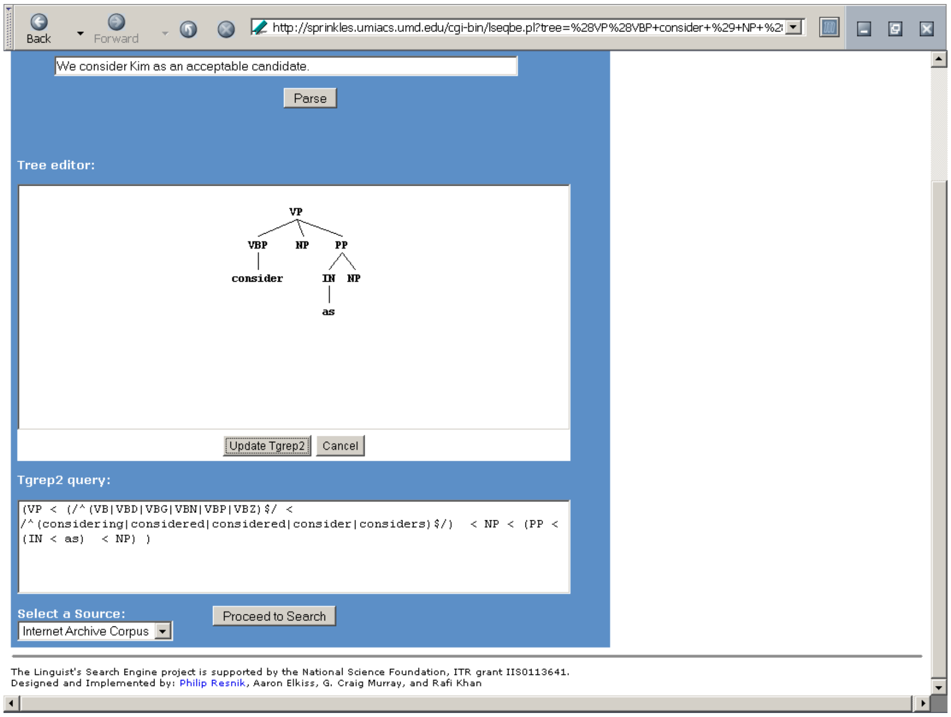
Once having logged in to the LSE, **you will find yourself in (or can easily go to) the Query By Example** (QBE) page. This is designed to make it easy for a linguist to say "Find me more examples like this one" *without* having to know the syntactic details underlying the LSE's annotations. The LSE currently uses a rather "vanilla" style of syntactic constituency annotation (of the Penn Treebank variety).

Type the sentence "We consider Kim as an acceptable candidate" into the *Example Sentence* space, and then click *Parse*. After a moment, you should see a parse tree for the sentence show up in the *Tree Editor* space.

Right-click on the VP node in the parse tree. This will bring up a menu of tree-editing operations. **Select *Remove all but subtree*.** You will see the tree display change so that only the VP subtree remains – we're interested in sentences containing this VP structure but we don't care about what's in the subject position, or whether or not it's a matrix sentence.

Right-click on the NNP above *Kim* to bring up the same menu. This time, **select *Remove subtree***. This will leave the NP dominated by VP, removing the unnecessary detail below – we care that the VP have an NP argument, but not what that NP contains.

Repeat the above *remove subtree* operation for each of DT, JJ, and NN. (At some point soon, we will probably add a *remove children* menu item to make it easier to remove all the children of a node at once.)



We consider Kim as an acceptable candidate.

```
(VP < (/^(VB|VBD|VBG|VBN|VBP|VBZ)$/ <
/^(considering|considered|considered|consider|considers)$/)  < NP < (PP <
(IN < as)  < NP) )
```

Internet Archive Corpus

The Linguist's Search Engine project is supported by the National Science Foundation, ITR grant IIS0113641.
Designed and Implemented by: Philip Resnik, Aaron Elkiss, G. Craig Murray, and Rafi Khan

At this point, your tree should look like the tree in the screen above. You have specified that you want verb phrases headed by *consider* where the VP also dominates an NP and a PP headed by *as*.

Now click the *Update Tgrep2* button. This automatically (re-)generates a query based on the tree structure you have specified.[3]

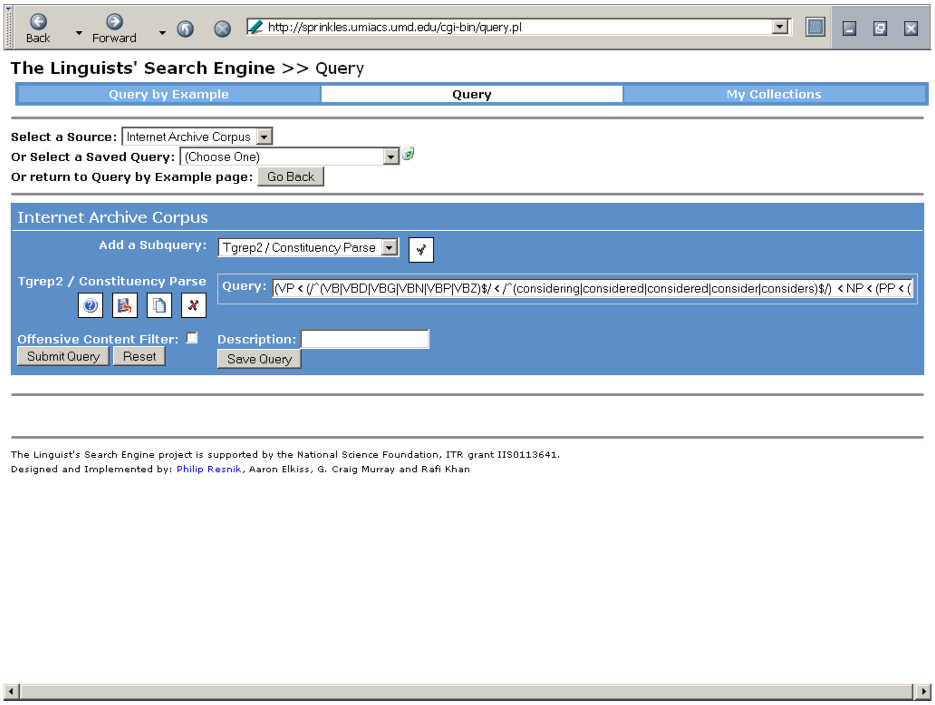
The screen above shows the resulting query in the *Tgrep2 query* area. The less-than sign (<) encodes the "immediately dominates" relation; e.g., part of the pattern says that there must be a node labeled with the nonterminal *IN* (Penntreebankese for *preposition*) that immediately dominates a node labeled with the word *as*. Notice that the LSE automatically expanded the tree-based pattern to include all grammatical inflections of the verb, not just present-tense *consider*. If there had been a lexical noun present, it would have included both the singular and plural forms. (For future versions of the LSE, we plan to extend the representation to include feature-based specifications, including not only tense and number features, but also semantic features such as WordNet class membership, Levin (1993) categories for verbs, etc.)

Advanced users can edit the *tgrep2* query here or in the screen that follows. See the "Tips, Hints, and Advanced Features" section for a detailed example.

Click *Proceed to Search* to move from *Query by Example* to the main search interface.

[3] The query language *tgrep2* is a variation of Rich Pito's original *tgrep*, distributed with the Penn Treebank. The *tgrep* family of tools lets you specify tree-based patterns to match in a parsed corpus (Rohde, 2001; http://tedlab.mit.edu/~dr/Tgrep2/).

The Query Interface



http://sprinkles.umiacs.umd.edu/cgi-bin/query.pl

The Linguists' Search Engine >> Query

Select a Source: Internet Archive Corpus

Or Select a Saved Query: (Choose One)

Or return to Query by Example page:

Internet Archive Corpus

Add a Subquery: Tgrep2 / Constituency Parse

Tgrep2 / Constituency Parse

Query: (VP < (/^(VB|VBD|VBG|VBN|VBP|VBZ)$/ < /^(considering|considered|considered|consider|considers)$/) < NP < (PP < (

Offensive Content Filter: □

Description:

The Linguist's Search Engine project is supported by the National Science Foundation, ITR grant IIS0113641.
Designed and Implemented by: Philip Resnik, Aaron Elkiss, G. Craig Murray and Rafi Khan

Let's look at the Query screen from top to bottom focusing on the most important pieces.

At the top, *Select a Source* allows you to choose what collection of sentences to look in. The default is currently a collection of several hundred thousand sentences collected from Web pages that are stored on the Internet Archive (www.archive.org). This static resource is a useful starting point for exploration; a little later you'll be shown how to create for yourself new collections of sentences from the Web that are likely to be of interest to you. Leave the source set to the Internet Archive Collection for now.

The *Select a Saved Query* pull-down allows you to recall queries that you've saved using the *Save Query* button at the bottom. This can be useful for modifying previous queries, or for trying out a query on a new source of sentences. Leave this alone for the moment, since we want to execute the query just created via Query By Example.

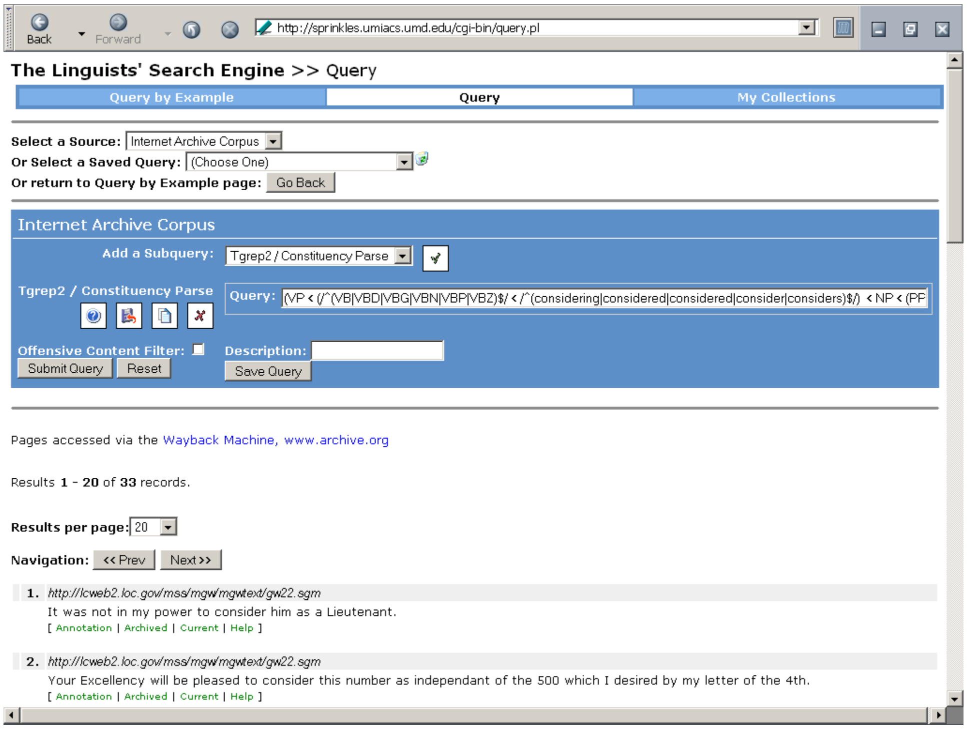
In the blue box are the search options when searching the collection of sentences from the Internet Archive. As we noted above, the query (*Tgrep2/Constituency Parse*) is expressed in terms of constituency (i.e. phrase structure) relationships.[4] To the left are a number of buttons we needn't deal with for the moment. You can click the *Offensive Content Filter* check box to apply a simple filter that will suppress URLs and sentences likely to be offensive.[5]

In the *Description* box at the bottom, type "consider NP as NP" and then click *Save Query*. This saves the query with a readable description to retrieve it by. **Then click *Submit Query***.

[4] Note for advanced users: these tree search expressions are *tgrep2* patterns. Advanced users could go directly to this page and type in arbitrary *tgrep2* queries rather than having Query By Example generate a valid pattern for you automatically. Also, the *Add a Subquery* button allows advanced users to specify secondary filtering criteria, e.g. more *tgrep2* patterns that must match. Sentences must match all subqueries to be returned, i.e. the subqueries are combined via Boolean *AND*.

[5] The Offensive Content Filter is based on a simple word-list approach – imagine George Carlin's list of "seven words you can't say on TV" expanded a great deal based on the sorts of things likely to show up on Web pornography sites. Please be aware that the filter is not perfect.

Looking at Results Returned by a Query



Select a Source: Internet Archive Corpus
Or Select a Saved Query: (Choose One)
Or return to Query by Example page: Go Back

Internet Archive Corpus

Add a Subquery: Tgrep2 / Constituency Parse

Tgrep2 / Constituency Parse **Query:** (VP < (/^(VB|VBD|VBG|VBN|VBP|VBZ)$/ < /^(considering|considered|considered|consider|considers)$/) < NP < (PP

Offensive Content Filter: □ **Description:**

Pages accessed via the Wayback Machine, www.archive.org

Results **1 - 20** of **33** records.

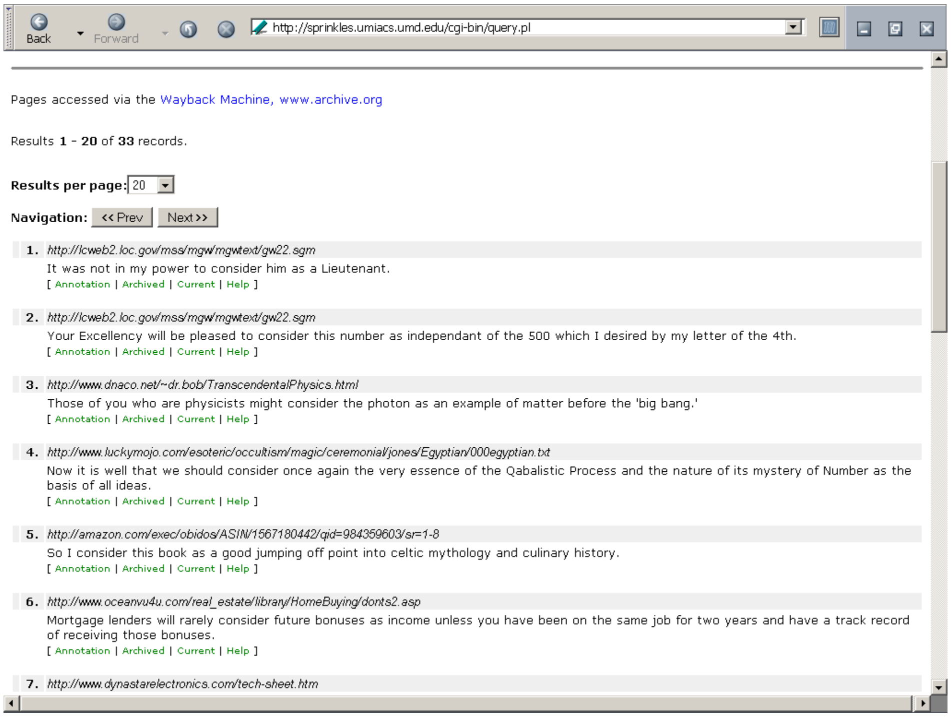
Results per page: 20

Navigation: << Prev Next >>

1. *http://lcweb2.loc.gov/mss/mgw/mgwtext/gw22.sgm*
It was not in my power to consider him as a Lieutenant.
[Annotation | Archived | Current | Help]

2. *http://lcweb2.loc.gov/mss/mgw/mgwtext/gw22.sgm*
Your Excellency will be pleased to consider this number as independant of the 500 which I desired by my letter of the 4th.
[Annotation | Archived | Current | Help]

The screen above shows results of your query. Notice that the "hits" are organized in standard search engine fashion, showing the number of matching sentences found, the URL of the page where each sentence was found, the sentence itself, navigation buttons to get to the next and previous twenty hits, etc. **Scroll down to get the view below, showing the first six hits.**



Pages accessed via the Wayback Machine, www.archive.org

Results **1 - 20** of **33** records.

Results per page: 20

Navigation: << Prev Next >>

1. *http://lcweb2.loc.gov/mss/mgw/mgwtext/gw22.sgm*
It was not in my power to consider him as a Lieutenant.
[Annotation | Archived | Current | Help]

2. *http://lcweb2.loc.gov/mss/mgw/mgwtext/gw22.sgm*
Your Excellency will be pleased to consider this number as independant of the 500 which I desired by my letter of the 4th.
[Annotation | Archived | Current | Help]

3. *http://www.dnaco.net/~dr.bob/TranscendentalPhysics.html*
Those of you who are physicists might consider the photon as an example of matter before the 'big bang.'
[Annotation | Archived | Current | Help]

4. *http://www.luckymojo.com/esoteric/occultism/magic/ceremonial/jones/Egyptian/000egyptian.txt*
Now it is well that we should consider once again the very essence of the Qabalistic Process and the nature of its mystery of Number as the basis of all ideas.
[Annotation | Archived | Current | Help]

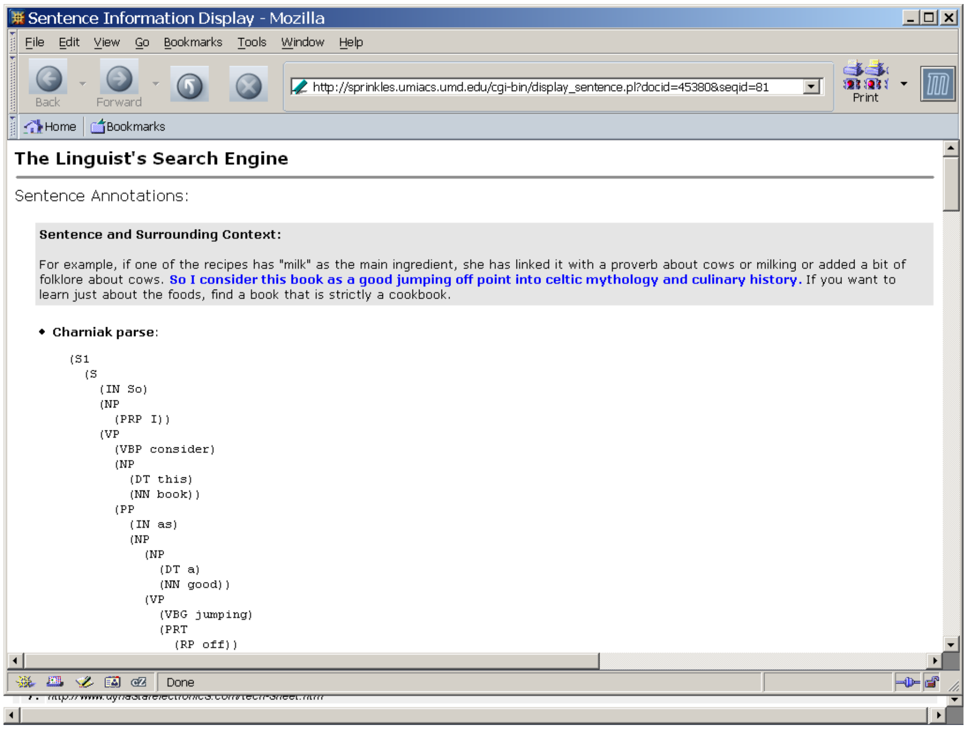
5. *http://amazon.com/exec/obidos/ASIN/1567180442/qid=984359603/sr=1-8*
So I consider this book as a good jumping off point into celtic mythology and culinary history.
[Annotation | Archived | Current | Help]

6. *http://www.oceanvu4u.com/real_estate/library/HomeBuying/donts2.asp*
Mortgage lenders will rarely consider future bonuses as income unless you have been on the same job for two years and have a track record of receiving those bonuses.
[Annotation | Archived | Current | Help]

7. *http://www.dynastarelectronics.com/tech-sheet.htm*

Notice that some hits, like the first one, are using “consider NP as NP” in the wrong way, e.g. “consider NP as a *candidate* for NP”. But hit number 5 looks like it’s probably a counterexample to the claim in (1e).

Click the *Annotation* link below hit number 5. This will bring you to a screen like this one.


Sentence Information Display - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back
Forward
Print
Home
Bookmarks
Done

http://sprinkles.umiacs.umd.edu/cgi-bin/display_sentence.pl?docid=45380&seqid=81

The Linguist's Search Engine

Sentence Annotations:

Sentence and Surrounding Context:

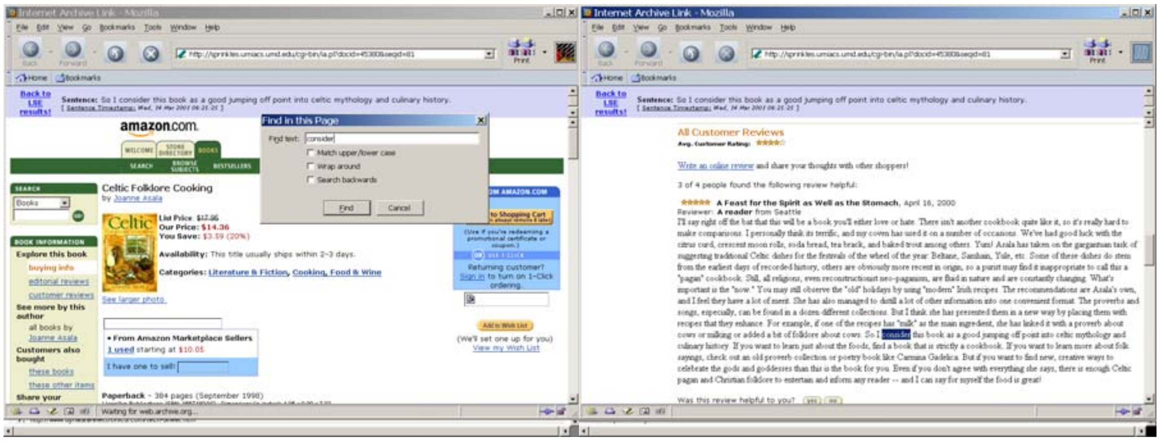
For example, if one of the recipes has "milk" as the main ingredient, she has linked it with a proverb about cows or milking or added a bit of folklore about cows. **So I consider this book as a good jumping off point into celtic mythology and culinary history.** If you want to learn just about the foods, find a book that is strictly a cookbook.

- **Charniak parse:**

```
(S1
  (S
    (IN So)
    (NP
      (PRP I))
    (VP
      (VBP consider)
      (NP
        (DT this)
        (NN book))
      (PP
        (IN as)
        (NP
          (NP
            (DT a)
            (NN good))
          (VP
            (VBG jumping)
            (PRT
              (RP off))
```

Notice that this shows the previous and following sentence context, and a number of linguistic annotations of the sentence, including, for example, the constituency parse. Scroll down to look at the full set of annotations. Then **go back to the list of hits**.

Next, **click on the *Archived* link**. This brings you to the Web page containing the sentence, as stored on the Internet Archive:



http://sprinkles.umiacs.umd.edu/cgi-bin/la.pl?docid=45380&seqid=81

Sentence: So I consider this book as a good jumping off point into celtic mythology and culinary history.
[Sentence Timestamp: Wed, 14 Mar 2007 06:25:25]

Celtic Folklore Cooking
by Joanne Asala

List Price: ~~$17.95~~
Our Price: $14.36
You Save: $3.59 (20%)

Availability: This title usually ships within 2-3 days.

Categories: Literature & Fiction, Cooking, Food & Wine

• From Amazon Marketplace Sellers
1 used starting at $10.05

I have one to sell!

Paperback - 384 pages (September 1998)

Find text: consider
□ Match upper/lower case
□ Wrap around
□ Search backwards

(Use if you're redeeming a promotional certificate or coupon.)

All Customer Reviews
Avg. Customer Rating: ★★★★☆

3 of 4 people found the following review helpful:

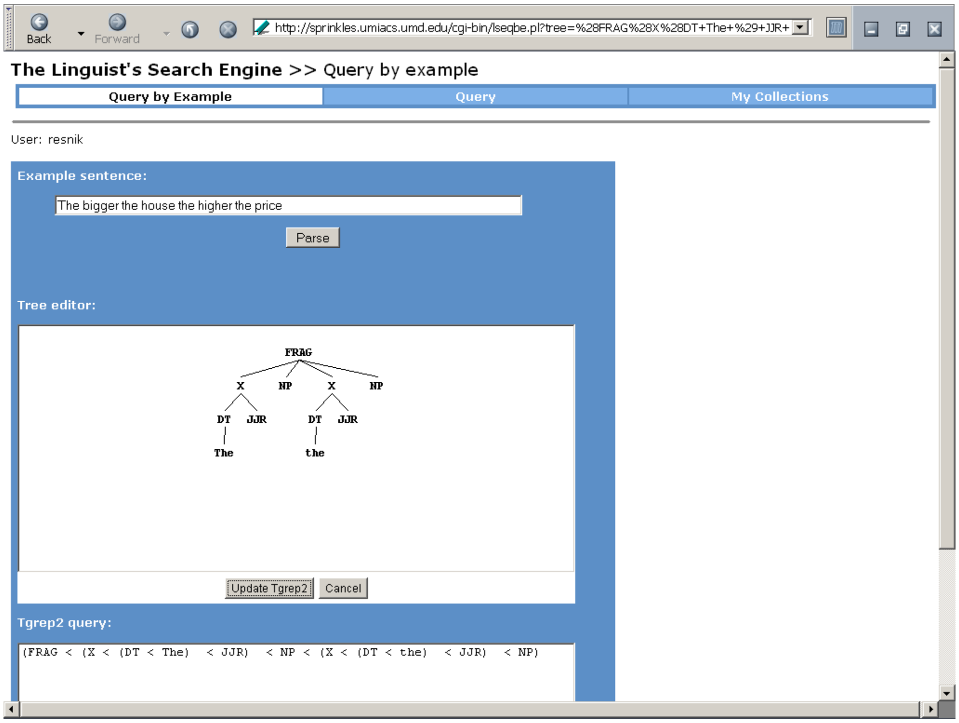
★★★★★ **A Feast for the Spirit as Well as the Stomach**, April 16, 2000
Reviewer: **A reader** from Seattle
I'll say right off the bat that this will be a book you'll either love or hate. There isn't another cookbook quite like it, so it's really hard to make comparisons. I personally think its terrific, and my coven has used it on a number of occasions. We've had good luck with the citrus curd, crescent moon rolls, soda bread, tea brack, and baked trout among others. Yum! Asala has taken on the gargantuan task of suggesting traditional Celtic dishes for the festivals of the wheel of the year: Beltane, Samhain, Yule, etc. Some of these dishes do stem from the earliest days of recorded history, others are obviously more recent in origin, so a purist may find it inappropriate to call this a "pagan" cookbook. Still, all religions, even reconstructionist neo-paganism, are fluid in nature and are constantly changing. What's important is the "now." You may still observe the "old" holidays by using "modern" Irish recipes. The recommendations are Asala's own, and I feel they have a lot of merit. She has also managed to distill a lot of other information into one convenient format. The proverbs and songs, especially, can be found in a dozen different collections. But I think she has presented them in a new way by placing them with recipes that they enhance. For example, if one of the recipes has "milk" as the main ingredient, she has linked it with a proverb about cows or milking or added a bit of folklore about cows. So I consider this book as a good jumping off point into celtic mythology and culinary history. If you want to learn just about the foods, find a book that is strictly a cookbook. If you want to learn more about folk sayings, check out an old proverb collection or poetry book like Carmina Gadelica. But if you want to find new, creative ways to celebrate the gods and goddesses than this is the book for you. Even if you don't agree with everything she says, there is enough Celtic pagan and Christian folklore to entertain and inform any reader -- and I can say for myself the food is great!

You can **use your Web browser’s “Find” function to find the sentence on the page**. You can **go back and click *Current* to see the current version of the page**, which may have changed (and therefore may or may not still contain the sentence).

Another Query by Example

Let's try another Query by Example. This time we'll look for instances of a construction (Goldberg, 1995) – in this case sentences containing things like "the ADJer the NP the ADJer the NP". **Go back to the *Query by Example* page** (you can click on it on the navigation bar at the top or bottom of most LSE pages) and **type in, as the example sentence, "The bigger the house the higher the price"** (without the double quotes). Then **click *Parse***.

As an exercise**, use the tree editing functionality to modify the parse tree so it looks like the tree on the screen below**.



http://sprinkles.umiacs.umd.edu/cgi-bin/lseqbe.pl?tree=%28FRAG%28X%28DT+The+%29+JJR+

The Linguist's Search Engine >> Query by example

User: resnik

Example sentence:

The bigger the house the higher the price

Tree editor:

Tgrep2 query:

```
(FRAG < (X < (DT < The)  < JJR)  < NP < (X < (DT < the)  < JJR)  < NP)
```

Remember, you right click on nodes to do things with them. You can also right click on the white space in the tree editor. Notice that the right-click menu includes *Undo*, which will undo your last operation if you make a mistake. You can also select *Revert*, or click the *Cancel* button at the bottom, to revert back to what the tree looked like before you started editing it. If you use the *Add Node…* option, you'll get a pop-up box in which to type the label of the new node you're adding.

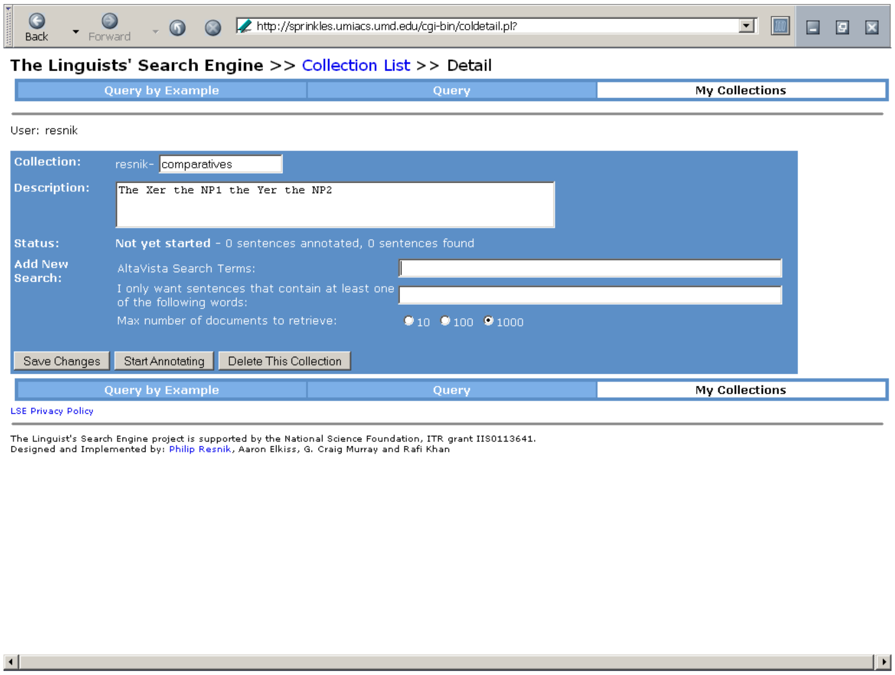
When your tree looks like the tree above, click *Update Tgrep2* to re-generate the query pattern. You'll have noticed that the parser really didn't know what to make of this construction. But that doesn't stop you from being able to edit the structure to generalize it (even if you don't know that JJR is the Penn Treebank symbol for comparative adjective), and it doesn't matter whether or not you *agree* with the structure as long as the resulting pattern can do a reasonable job of locating sentences with the *same* structure.

Click *Proceed to Search*.
Then enter the description **"The ADJer the NP the ADJer the NP"** and **click the *Save Query* button**.
Finally, **click *Submit Query***.

Uh oh… You'll notice that there were no matches for this query in the collection of Internet Archive sentences. But then, those sentences were collected randomly, and even if several hundred thousand sentences seems like a large number to search, it's a fairly small collection relative to the size of the Web. It's not surprising that any given construction might not appear in this particular random sample. What you *really* want to do is a Web-scale search, so that you can look for your structure on a *non*-random sample.

Building Your Own Collections

Let's use the LSE to do a large-scale Web search for instances of the "the ADJer the NP the ADJer the NP" construction. To start, **go to the *My Collections*** page and **click on *Add New Collection Definition***.



http://sprinkles.umiacs.umd.edu/cgi-bin/coldetail.pl?

The Linguists' Search Engine >> Collection List >> Detail

User: resnik

Collection: resnik- comparatives

Description: The Xer the NP1 the Yer the NP2

Status: **Not yet started** - 0 sentences annotated, 0 sentences found

Add New Search:

AltaVista Search Terms:

I only want sentences that contain at least one of the following words:

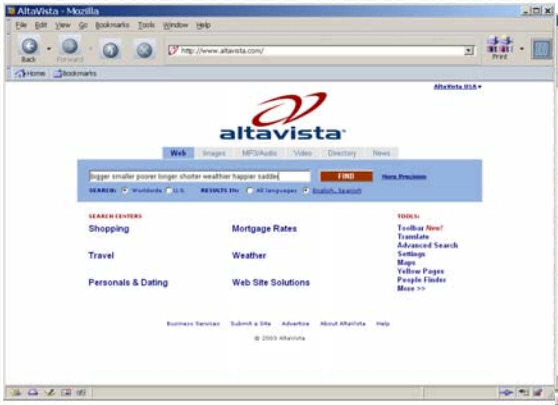
Max number of documents to retrieve: ○ 10 ○ 100 ● 1000

LSE Privacy Policy

The Linguist's Search Engine project is supported by the National Science Foundation, ITR grant IIS0113641.
Designed and Implemented by: Philip Resnik, Aaron Elkiss, G. Craig Murray and Rafi Khan

The *Collection* space allows you to give a descriptive name to this collection.[6] **Type "comparatives" into the *collection* box** as illustrated above. **In the *Description* area, type "The Xer the NP1 the Yer the NP2"** – this is a short prose description of the collection of sentences you're building from the Web.

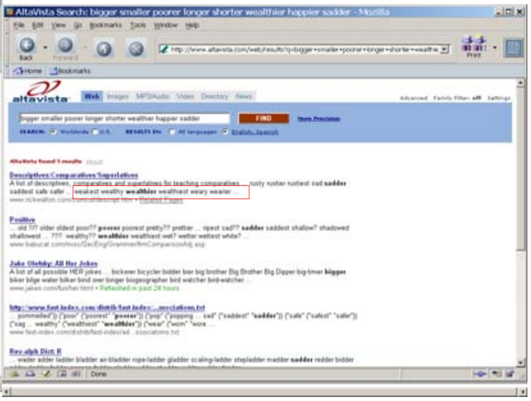
The *Add New Search* area is the heart of the collection building process. The key idea is (a) to use the Altavista search engine to find pages that are likely to contain sentences of interest, and then (b) to automatically extract those sentences of interest into a searchable LSE collection.



AltaVista - Mozilla

http://www.altavista.com/

bigger smaller poorer longer shorter wealthier happier sadder



AltaVista Search: bigger smaller poorer longer shorter wealthier happier sadder - Mozilla

bigger smaller poorer longer shorter wealthier happier sadder

Descriptives Comparatives Superlatives
A list of descriptives, comparatives and superlatives for teaching comparatives ... rusty rustier rustiest sad **sadder** saddest safe safer ... weakest wealthy **wealthier** wealthiest weary wearier ...

Positive
... old ??? older oldest poor?? **poorer** poorest pretty?? prettier ... ripest sad?? **sadder** saddest shallow? shadowed shallowest ... ???. wealthy?? **wealthier** wealthiest wet? wetter wettest white? ...

Jake Olefsky: All Her Jokes
A list of all possible HER jokes ... bickerer bicycler bidder bier big brother Big Brother Big Dipper big-timer **bigger** biker bilge water bilker bind over binger biogeographer bird watcher bird-watcher ...

Rev.alph Dict R
... wader adder ladder bladder air-bladder rope-ladder gladder scaling-ladder stepladder madder **sadder** redder bidder

The first step is done by opening a new browser window and using Altavista (www.av.com) to search for pages that are likely to contain sentences of interest. This can take a few iterations; for example, the screens above show that simply entering "bigger smaller longer poorer…etc." as an Altavista query won't work – it results in pages that contain word lists, rather than pages where those words are used in sentences.

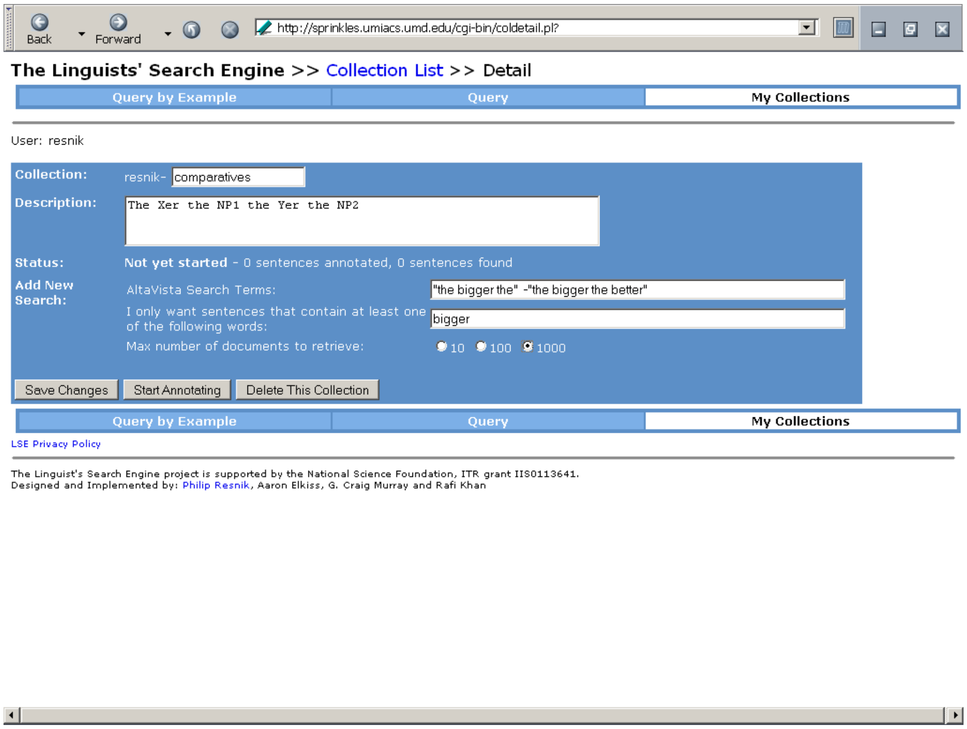
[6] For internal bookkeeping, collection names are always prefixed by the user's login name.

Here's an illustration of how to refine your Altavista query. **Go to Altavista and in the query box type `"the bigger the"`.** *Include* the double quotes, which tells Altavista you're interested in these three words appearing next to each other. You'll find that this gets you a lot of pages containing "the bigger the better", because it's such a common phrase. You can tell Altavista to exclude pages containing that phrase by adding a query term with a minus sign in front of it. **Type in this Altavista query**:

```
"the bigger the"  -"the bigger the better"
```

It says: get me pages containing "the bigger the" but not containing "the bigger the better". **Submit the query to Altavista** and notice that the hits you get back are indeed pages containing the right sorts of phrases.

Copy this query from your Altavista page, and then **go back to the LSE's *My Collections* screen**.



The Linguists' Search Engine >> Collection List >> Detail

User: resnik

Collection: resnik- comparatives

Description: The Xer the NP1 the Yer the NP2

Status: Not yet started - 0 sentences annotated, 0 sentences found

Add New Search:

AltaVista Search Terms: "the bigger the" -"the bigger the better"

I only want sentences that contain at least one of the following words: bigger

Max number of documents to retrieve: ○ 10 ○ 100 ● 1000

The Linguist's Search Engine project is supported by the National Science Foundation, ITR grant IIS0113641.
Designed and Implemented by: Philip Resnik, Aaron Elkiss, G. Craig Murray and Rafi Khan

Paste or type the query you copied into the LSE's *Altavista Search Terms* space. You've now told the LSE that it should automatically retrieve Web pages from Altavista using this query. The *Max number of documents to retrieve* defaults to 1000, though you can select a smaller number for testing purposes.

Since the Web pages you retrieve will undoubtedly contain many (mostly) sentences you are *not* interested in, there needs to be some way to specify which sentences you *are* interested in. The box underneath the Altavista search terms allows you to specify a word or words that must appear in a sentence in order for it to be interesting. **In the box saying *I only want sentences that contain at least one of the following words*, type "bigger"** (without quotes).

Now click *Save Changes*. You'll see that your collection now has a *Search 1* with the parameters you've given it.

Add new searches to this collection description by repeating the process above:

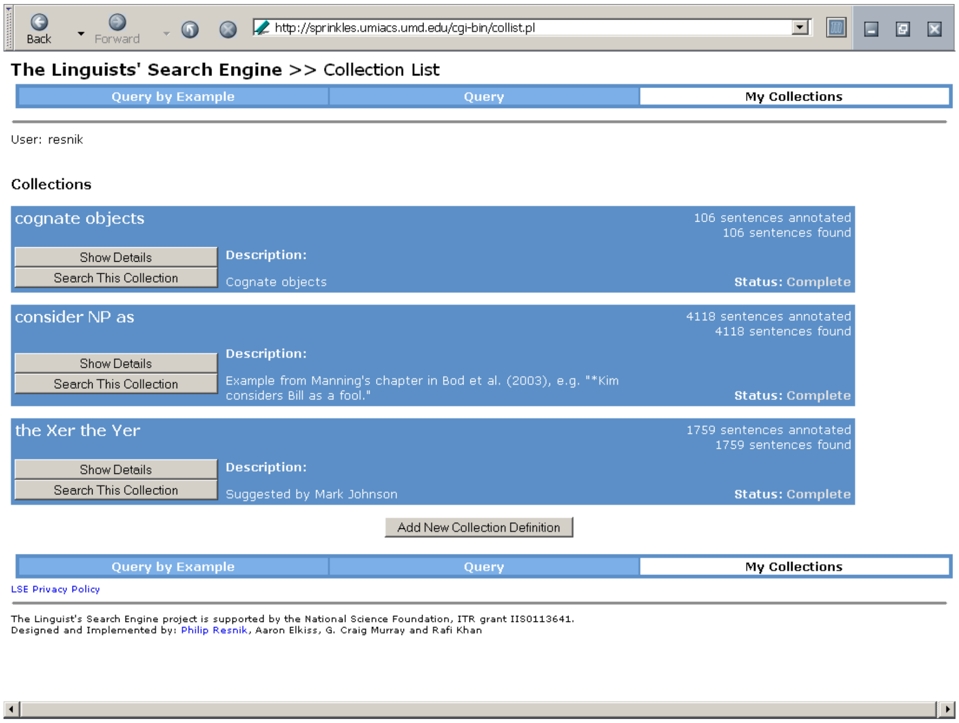
- (Optionally) Verify that your Altavista search retrieves the right sorts of pages
- Enter the Altavista search terms
- Enter the words that identify sentences of interest
- Choose the maximum number of documents to retrieve for this search
- Click the *Save Changes* button.

For example,

- Altavista search terms: `"the wealthier the"` *(include quotes)*
- I only want sentences…: `wealthier`
- Maximum number of documents: `1000`
- Click *Save Changes*

- Altavista search terms: `"the poorer the"` *(include quotes)*
- I only want sentences…: `poorer`
- Maximum number of documents: `1000`
- Click *Save Changes*

Go back to the *My Collections* page. Notice that this collection now appears on your list of collections. In the lower right corner, the *Status* line shows the current status of a collection. Possible values include *not yet started*, *queued* (i.e. waiting until the LSE annotator is free to work on it), *building/annotating*, and *complete*. Once the building and annotating process has started, sentences that are found are annotated as quickly as the LSE can get to them, given its available resources. Note that a collection is searchable as soon as it contains *any* annotated sentences, i.e. you don't have to wait for it to be complete.



The Linguists' Search Engine >> Collection List

User: resnik

Collections

cognate objects
106 sentences annotated
106 sentences found
Description:
Cognate objects
Status: Complete

consider NP as
4118 sentences annotated
4118 sentences found
Description:
Example from Manning's chapter in Bod et al. (2003), e.g. "*Kim considers Bill as a fool."
Status: Complete

the Xer the Yer
1759 sentences annotated
1759 sentences found
Description:
Suggested by Mark Johnson
Status: Complete

The Linguist's Search Engine project is supported by the National Science Foundation, ITR grant IIS0113641.
Designed and Implemented by: Philip Resnik, Aaron Elkiss, G. Craig Murray and Rafi Khan

At any point, you can click *Show Details* for a collection – for example, you can go back there to delete the collection, or to tell the LSE to stop annotating if the build is still in progress but you've already found everything you wanted. You can even add a new search to extend a collection that already exists.

Using Your Collections

The amount of time it takes to build a collection can vary – you can watch the *My Collections* list to see how things are progressing. It will show you how many sentences have been found so far that meet your criteria, and it will also show you how many of those have been linguistically annotated and are therefore now searchable.

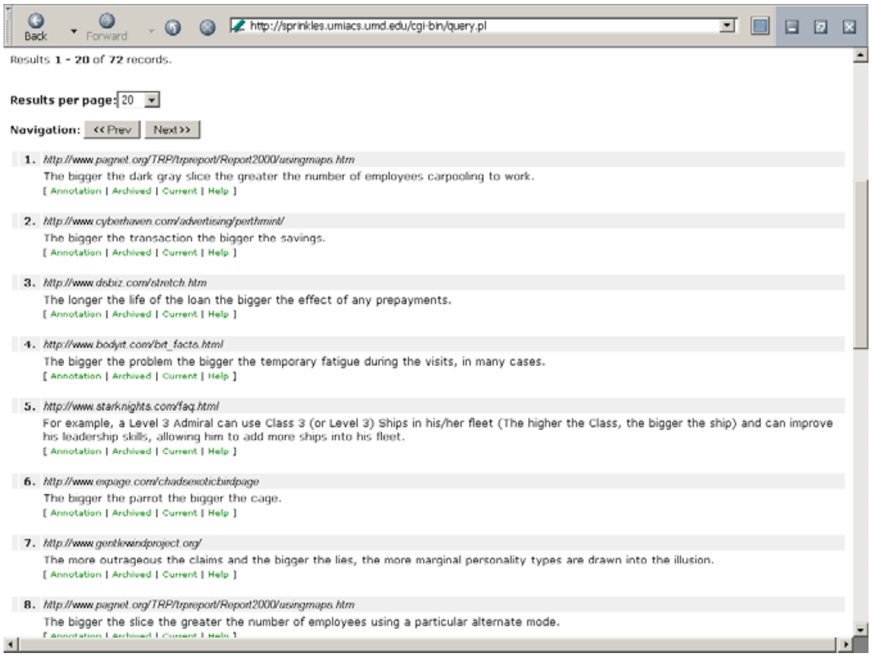
The LSE rotates its efforts among the requests of its various users, so your collection building request will *not* need to wait in line behind all the other requests in order for it to get started. Currently, the LSE's scheduler places a high priority on quickly getting some sentences into each collection – the first thousand – so that you can very quickly start searching and discover changes you need to make. (To conserve resources, please use *Delete Collection* for collections you've decided not to use, and use the *Stop Annotating* button for collections once they've grown as large as you need them.) After the first thousand sentences, you may notice that your collection builds up more slowly if other users are also building collections at the same time. The scheduler also keeps track of which collections have not received any attention for a while, to make sure that each one gets its fair share.

Let's return to the search for "The Xer the NP1 the Yer the NP2" constructions, using the collection you have built. (Remember, you can do this even before the collection is complete.)

Go back to the *Query* page. At the top, **use the pull-down for *Select a Source* to pick *Altavista Corpora***. Notice that the blue box now offers you a new option: **use the pull-down menu for *Choose Altavista Collection* to select the *comparatives* collection**.

Now use the *Select a Saved Query* pull-down at the top of the screen to pick the query you saved before: "The ADJer the NP the ADJer the NP". Notice that the LSE automatically fills in the query pattern for you.

Click *Submit Query*. Depending on how far your collection building has gotten, the results should look something like this:



Results **1 - 20** of **72** records.

Results per page: 20

Navigation: << Prev Next >>

1. *http://www.pagnet.org/TRP/trpreport/Report2000/usingmaps.htm*
   The bigger the dark gray slice the greater the number of employees carpooling to work.
   [Annotation | Archived | Current | Help]

2. *http://www.cyberhaven.com/advertising/perthmint/*
   The bigger the transaction the bigger the savings.
   [Annotation | Archived | Current | Help]

3. *http://www.dsbiz.com/stretch.htm*
   The longer the life of the loan the bigger the effect of any prepayments.
   [Annotation | Archived | Current | Help]

4. *http://www.bodyrt.com/brt_facts.html*
   The bigger the problem the bigger the temporary fatigue during the visits, in many cases.
   [Annotation | Archived | Current | Help]

5. *http://www.starknights.com/faq.html*
   For example, a Level 3 Admiral can use Class 3 (or Level 3) Ships in his/her fleet (The higher the Class, the bigger the ship) and can improve his leadership skills, allowing him to add more ships into his fleet.
   [Annotation | Archived | Current | Help]

6. *http://www.expage.com/chadsexoticbirdpage*
   The bigger the parrot the bigger the cage.
   [Annotation | Archived | Current | Help]

7. *http://www.gentlewindproject.org/*
   The more outrageous the claims and the bigger the lies, the more marginal personality types are drawn into the illusion.
   [Annotation | Archived | Current | Help]

8. *http://www.pagnet.org/TRP/trpreport/Report2000/usingmaps.htm*
   The bigger the slice the greater the number of employees using a particular alternate mode.
   [Annotation | Archived | Current | Help]

Congratulations! You have just searched the entire Web (or at least the portion indexed by Altavista) using a structural search, and found some examples of the structure you were looking for.

Tips, Hints, and Advanced Features

The examples above have exercised all of the LSE's basic functionality as of this writing. Here are few things that may help make it more useful, based on our experience so far:

- **Navigation bar.** The navigation bar at the top and bottom of most screens makes it easy to jump back and forth between Query by Example, Query, and My Collections.

- ***Search this Collection* shortcut.** When you're in My Collections, either in the collections list or in the detailed view of a particular collection, you can click *Search this Collection* to go to a version of the Query page where the collection information has already been filled in.

- **Tree editing hints.** Unless you are particularly interested in your structure's occurring at the matrix level, the usual first step will be to right-click on the deepest relevant node and select *Remove all but subtree*. If you're looking at a verb-centered construction and you don't need a matrix sentence (and the sentential subject doesn't matter), it's usually better to keep just the VP rather than the whole S dominating it, since Treebank-style parses will occasionally used adjoined structures (VP dominating VP). On the same note, we recommend being more general rather than more specific where possible – for example, unless you specifically need a particular NP-internal structure, we recommend keeping just the NP (as was done in the earlier examples) rather than, say, using a specification that requires a determiner. It's always easier to go from more general to more specific once you've seen what the data look like.

- **Excluding structure.** There are a few simple things you can do to the automatically generated *tgrep2* expression, without learning the whole complicated pattern-matching syntax, that are very useful; foremost among these is negation. The LSE's current Query By Example does not provide a way to say that a part of a structure should be *absent* rather than present; for example, the tree editor does not allow you to say that an NP should not contain an adjective, or that a VP should not have a PP as one of its children. One way to get this behavior is to type in an example sentence that *includes* the structure you don't want, generate the *tgrep2* expression automatically, and then modify it manually to negate the relevant piece of structure. For example, suppose you want cognate object constructions for the verb *live* where the direct object does not have an adjectival modifier ("lived a/the/his/her life", but not "lived a quiet life").

    - In Query by Example, type "He lived a quiet life", click *Parse*, and edit the tree to keep just the VP. Use *remove subtree* to delete the DT (determiner) node, but keep the JJ (adjective) subtree. Click *Update Tgrep2*.

    - In the *tgrep2* query, scroll right, if necessary, so you can see the part of the pattern that specifies the object noun phrase:

      ```
      (NP < (JJ < quiet) < (/^(NN|NNS)$/ …etc.
      ```

      The first greater-than sign stands for immediate dominance, so this says that we want an NP that dominates a JJ node (which itself dominates a node labeled *quiet*), and that also dominates a subtree where the root node is labeled NN or NNS, etc. If you put an exclamation point before the greater-than sign (`!<`) you change it from *dominates* to *does not dominate*, so if you changed the expression this way

      ```
      (NP < (JJ !< quiet) < (/^(NN|NNS)$/ …etc.
      ```

      then you have modified your structure to specify an NP that must contain an adjective (JJ), but you've said that that adjective cannot be the word *quiet*. And, in fact, you could say

      ```
      (NP < (JJ !< quiet|peaceful|good) < (/^(NN|NNS)$/ …etc.
      ```

in order to exclude the adjectives *quiet*, *peaceful*, and *good* (the vertical bar means “or”).

This, however, is not quite want we wanted – we wanted to exclude *all* adjectives. The way to do this is to change the specification so that the negation applies to the whole JJ (adjective) and doesn’t care about what’s underneath it:

```
(NP !< JJ < (/^(NN|NNS)$/ …etc.
```

  - Once you’ve edited the query, you can *Proceed to Search*, save the query, etc., as usual. (Note you can edit the *tgrep2* expression on the query page, as well.) If you execute this query in the Internet Archive collection of sentences, you’ll get sentences like “You might get hurt, but it's the only way to live life completely”, etc.

  - **Exercise**: If you wanted to specify a live-life cognate object construction *with no post-verbal adverb*, i.e. excluding the above sentence, what *tgrep2* expression would you come up with? See footnote for one answer.[7]

- **If the LSE gets stuck.** If the LSE gets into a strange state that you can’t get out of, the first thing to try is using your browser to force a reload of the page (in most browsers, hold *shift* and click the *reload page* button). The second thing to try is navigating off the page and then navigating back to it, again perhaps reloading it when you get there. The third thing to try is quitting out of your browser entirely, and then starting up the browser again and going to the LSE. As with all things computational, *save frequently* (e.g. using the *Save Query* button) if there’s something that’s important.

- **Logging out.** There is currently no functionality for logging out. You can just quit your browser.

- **Use the LSE discussion group.** A Yahoo group has been set up for LSE users, called lse_support. Join the group, help each other out, and above all please give us feedback on ways to improve the LSE and which features are most important to add next.

- **Have fun, do good work, and keep us posted!** The future of the LSE depends, in part, on whether or not it turns out to support good linguistics research. We would very much like to keep track of presentations, papers, articles, and projects where the LSE has played a role.

[7] Using Query by Example with the sentence “It’s the only way to live life completely” and editing the tree and the pattern as recommended, you can get to the expression `(VP < (/^(VB|VBD|VBG|VBN|VBP|VBZ)$/ < /^(lived|lived|lives|living|live)$/)  < (NP < (/^(NN|NNS)$/ < /^(lives|lives's|life's|life)$/) )  !< ADVP).` Crucially, notice the exclamation point near the end of the expression, which is saying that the VP should *not* dominate an ADVP.

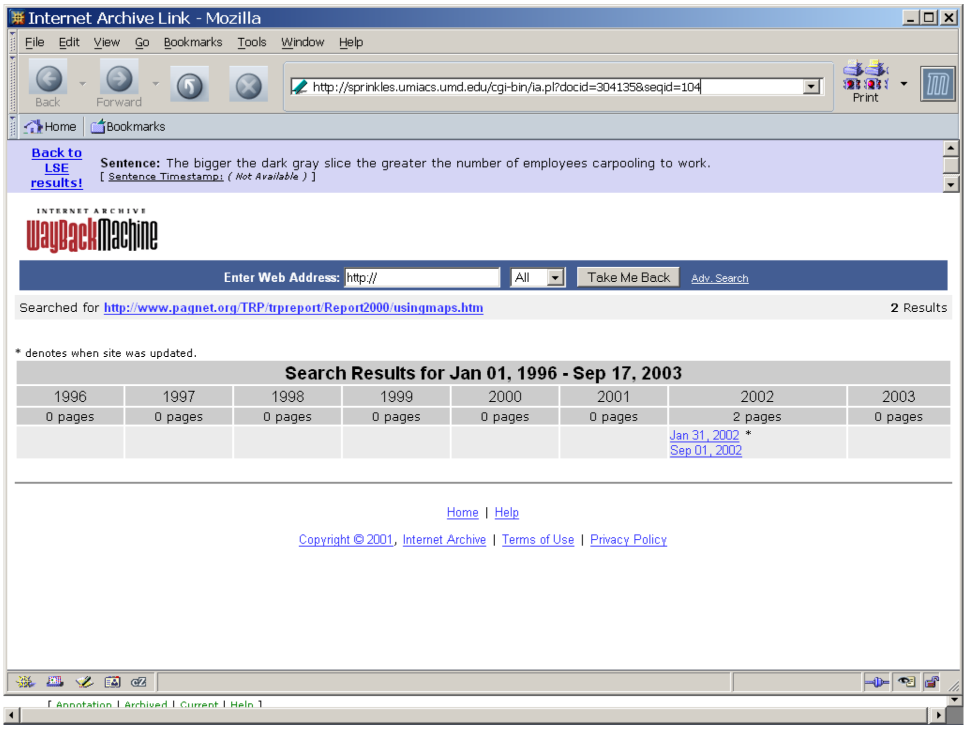
Appendix: Citing Data Found Using the LSE

In presentations and publications using Web data, we strongly recommend careful documentation of the sources of those data – not only as good research practice, but to bolster the credibility of the data, since anyone who doubts a claim ("Are you sure that sentence came from a page where the person really knew English?") can go to the data and decide for himself or herself.

The Internet Archive collection makes this particularly easy: for sentences found in this collection, we recommend providing the Internet Archive's URL for the page, which includes the page's original URL plus a timestamp identifying the date the page was crawled.[8]

It's worth noting that, unlike the collection of Internet Archive sentences, Altavista collection sentences are taken from current pages on the Web, which might change or cease to exist at any time. This is undesirable in terms of having persistent data that anyone can return to, but a minimum, the APA style guide recommends that, "a reference of an Internet source should provide a document title or description, a date (either the date of publication or update or the date of retrieval), and an address (in Internet terms, a uniform resource locator, or URL)" (http://www.apastyle.org/elecgeneral.html, retrieved 4 October 2003).

For pages in Altavista-based collections, the LSE will help you find a more permanent citation by making it easy to locate stored snapshots of this page on the Internet Archive. If you click the *Archived* link below a hit, for a sentence that came from an Altavista-based collection, the LSE will look on the Internet Archive and will show you its list of snapshots for that page.


Internet Archive Link - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back
Forward
Print
Home
Bookmarks
Back to LSE results!
INTERNET ARCHIVE
WaybackMachine
Enter Web Address:
All
Take Me Back
Adv. Search
Home | Help
Copyright © 2001, Internet Archive | Terms of Use | Privacy Policy
[Annotation | Archived | Current | Help]

http://sprinkles.umiacs.umd.edu/cgi-bin/ia.pl?docid=304135&seqid=104

Sentence: The bigger the dark gray slice the greater the number of employees carpooling to work.
[Sentence Timestamp: (*Not Available*)]

Enter Web Address: http://

Searched for http://www.pagnet.org/TRP/trpreport/Report2000/usingmaps.htm **2** Results

* denotes when site was updated.

Search Results for Jan 01, 1996 - Sep 17, 2003

1996	1997	1998	1999	2000	2001	2002	2003
0 pages	0 pages	0 pages	0 pages	0 pages	0 pages	2 pages	0 pages
						Jan 31, 2002 * Sep 01, 2002	

One of these snapshots may be a permanently archived version of the page that contains the sentence you're looking at. In our opinion, it is worth looking for the Internet Archive version of any data that you consider important.

[8] This will very shortly be added to the information available via a sentence's *Annotation* link.

Bibliography

Steven Abney, “Statistical Methods and Linguistics”, in J. Klavans and P. Resnik (eds.), *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, Cambridge, MA: MIT Press, pp. 1-26, 1996.

American National Corpus, http://americannationalcorpus.org/, as of 9 November 2003.

British National Corpus, http://www.natcorp.ox.ac.uk/, as of 9 November 2003.

Child Language Data Exchange System (CHILDES), http://childes.psy.cmu.edu/, as of 9 November 2003.

Corley, S., Corley, M., Keller, F., Crocker, M., & Trewin, S., “Finding Syntactic Structure in Unparsed Corpora: The Gsearch Corpus Query System”. *Computers and the Humanities*, 35, 81-94, 2001.

FrameNet, http://www.icsi.berkeley.edu/~framenet/, as of 9 November 2003.

Francis, S. and H. Kučera, Computing Analysis of Present-day American English, Brown University Press, Providence, RI, 1967.

Goldberg, Adele E. *Constructions: A Construction Grammar Approach to Argument Structure*, University of Chicago Press, 1995.

Levin, Beth, *English Verb Classes And Alternations: A Preliminary Investigation*, Chicago: University of Chicago Press, 1993.

Linguistic Data Consortium (LDC), http://www.ldc.upenn.edu/, as of 9 November 2003.

Manning, Christopher D. “Probabilistic Syntax”, in Rens Bod, Jennifer Hay, and Stefanie Jannedy (eds*), Probabilistic Linguistics*, pp. 289-341. Cambridge, MA: MIT Press, 2003.

Oostdijk, N. & P. de Haan (eds.). *Corpus-based research into language*. Amsterdam: Rodopi. 1994.

Penn Treebank, http://www.cis.upenn.edu/~treebank/home.html, as of 9 November 2003.

Pollard, C. and I. A. Sag, *Head-Driven Phrase Structre Grammar*. Chicago: University of Chicago Press, 1994.

PropBank, http://www.cis.upenn.edu/~ace/, as of 9 November 2003.

Rohde, D., *Tgrep2*, http://tedlab.mit.edu/~dr/Tgrep2/, 2001, page as of 9 November 2003.

Sapir, Edward. Language: *An Introduction to the Study of Speech*. New York: Harcourt, Brace, 1921; Bartleby.com, 2000. www.bartleby.com/186/.

Scott, M., *Wordsmith Tools version 3*, Oxford: Oxford University Press. ISBN 0-19-459289-8, 1999.