

The Linguist's Search Engine User's Guide

Aaron Elkiss and Philip Resnik

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

24-Aug-2004

1 Introduction

This document describes how to use the Linguist's Search Engine to perform syntactic searches on Internet data. Currently available are a three million sentence corpus of sentences from the Internet Archive as well as facilities to build and search corpora based around search results from AltaVista queries.

2 First Steps

2.1 Requirements

The Linguist's Search Engine makes extensive use of JavaScript, Cascading Style Sheets, and Java, so an up-to-date browser is a must.

The recommended browsers for the Linguist's Search Engine are Mozilla 1.5 or later (earlier versions have a bug with Java that can cause random crashes), Mozilla Firefox 0.7 or later and Internet Explorer 6.

You will also need to install the Sun Java Runtime Environment (JRE), which is available from <http://www.java.com> by clicking on the yellow "Get It Now" button.

2.2 Registration

Before you can use the Linguist's Search Engine, you must create an account by clicking on the **Register to use the Linguist's Search Engine** link and filling out the form.. This allows any collections you build to be marked as belonging to you (see 4)

The only required information is your name, email address, desired user id and password. Optionally, you can include your institution, title and research interests. This gives us an idea of who is using the Linguist's Search Engine.

When you submit the form, you will receive a confirmation message and a link back to the LSE front page. In addition, you will receive a message at the email address you specified confirming your registration.

If the passwords or email addresses provided do not match or if you fail to provide one of the required pieces of information, you will receive a message to that effect and be given the option to return to the registration page and enter the missing information.

2.3 Logging In

You should be able to log in as soon as you create an account - there is no need to wait to receive the confirmation email.

To log in, click the **Log in to the Linguist's Search Engine** link on the LSE front page. Enter the username and password you specified when registering.

A login cookie will be saved on your computer (it does not contain your username or your password), so you should click "Log Out" at the top of the Query or My Collections pages when you are done if you are using a public computer.

2.4 Forum

There is a forum available for questions and discussion about the Linguist's Search Engine. Select **Linguist's Search Engine Forum** from the front page. Help on using the forum is available by clicking **FAQ** from the forum's navigation menu.

3 Querying

The default page after logging in is the Query page. Here you can enter sample sentences, view and manipulate their parse trees, and execute, save and load

queries.

You can effectively create queries for the Linguist's Search Engine by parsing a sentence that contains the structure in which you are interested, removing all but that particular structure from the parse tree and searching for sentences whose parses contain your query as a subtree.

3.1 Language Selection

The Linguist's Search Engine supports collections in multiple languages. You can choose the language in which you want to enter queries and view results by selecting the language from the drop-down box in the "Query by Example" section of the page. Assuming your computer is set up to display text in the language you have selected, you should see a new parsed example sentence in the chosen language. You will also be able to enter and parse a new sentence in the selected language.

When you change the language, the collection to search (under Query Options) will be reset to the default for the given language. Also, only collections with parses in the selected language will be displayed as choices for the collection to search. Some collections (parsed parallel corpora) may support searching in multiple languages.

3.2 Parsing a Sentence

To parse and view a sentence, enter the sentence you wish to parse in the **Example Sentence** text box and click the **Parse** button. Your sentence will be automatically parsed and the tree will be displayed in the tree editor. A textual representation of the parse will be displayed in the **Query** text area. You can show and hide the example sentence input field by clicking on the + or - icon next to it..

Note that currently sentences longer than 20 words will not be parsed on the Query by Example page.

3.3 Tree Editor

The tree editor allows you to add, remove and edit nodes in the tree. You can show and hide the tree editor by clicking on the + or - icon next the tree editor.

There are five basic operations that can be performed in the tree editor: Edit Node, Undo, Revert, Add Node, Remove Nodes, and Change Relation.

These operations are invoked either by right-clicking on empty space in the tree editor (for Undo and Revert), or by right-clicking on a node in the parse tree

(for the other options).

You can also change how a word will be expanded. Double-clicking on a word brings up the expansion options menu. Here you can use WordNet to replace words with classes of words as well allow the query to match any morphological form of the word.

3.3.1 Edit Node

Right-clicking on a node and selecting “Edit Node” brings up a dialog box that allows you to enter a new label for the node. Enter the new label and click “OK”.

3.3.2 Undo

To undo the last operation performed in the tree editor, right-click anywhere in the tree editor and click **Undo**.

3.3.3 Revert

To revert the tree to the state it was in when the page was most recently loaded, right-click anywhere in the tree editor and click **Revert**.

3.3.4 Add Node

Nodes can be added in 4 places - as a child, left sibling, right sibling, and parent of a selected node.

To add a new node, left click on the node that will be the parent, child, or sibling of the node you wish to add. Right click on this node, select **Add Node**, and select the relationship the new node should have to the selected node.

If you add a node as a parent to a node that is not the root, the parent of the existing node will become the parent of the new node.

3.3.5 Remove

The remove submenu allows you to remove nodes from the tree.

Remove Subtree To remove a subtree rooted at a particular node, left click on the node that is the root of the subtree you wish to remove, right click, and select **Remove -> Subtree**.

Remove All but Subtree To remove all of the tree but a particular subtree of interest, left click on the node that is the root of the subtree you wish to keep, right click, and select **Remove -> All but Subtree**.

Remove All Children To remove all child nodes of a particular node, left click on the node that is the parent of the nodes you wish to remove, right click, and select **Remove -> All Children**. All the children of the selected node as well as all their descendants will be removed.

3.3.6 Relations

The relations submenu allows you to change the relation a node and its parent are required to have from its children in the query tree.

Immediately dominates subtree This is the default relation. The relation of the parent to the selected node and its descendants (the “subtree”) is required to be that of immediate dominance, that is, the subtree must appear exactly in that location in the query results.

Does not immediately dominate subtree Right-clicking on a node and selecting **Relations -> Does not immediately dominate subtree** will require the selected node and its descendants not to appear in query results in that location in the parse tree.

Dominates subtree Right-clicking on a node and selecting **Relations -> Dominates subtree** will relax the requirement of immediate dominance and allow the subtree (the selected node and its descendants) to appear in query results anywhere as a descendant of the parent node.

Does not dominate subtree Right-clicking on a node and selecting **Relations -> Dominates subtree** will require the subtree (the selected node and its descendants) *not* to appear in query results anywhere as a descendant of the parent node.

3.3.7 Node Expansion

Nodes in the query that are words can be expanded to a list of words. Currently, words can be expanded to all morphological forms of the word, the results of a WordNet relation (for example, replacing the word “eat” with particular ways

to eat, i.e. troponyms of “eat”), or all morphological forms of all results of a WordNet relation.

Only words can be directly expanded - they will expand their part of speech tags along with them. For example, selecting a noun tagged NN and choosing to use all word forms would result in the part of speech tag also allowing NNS. Phrase tags cannot be expanded.

To change the options for how a word is expanded, double click on the word. Change “use given form” to “use all word forms” and click “OK” to expand a verb or noun into all its morphological forms.

3.3.8 WordNet expansions

The tree editor can be used to expand tokens based on WordNet relationships. Currently we are using the WordNet 2.0 database.

To replace a noun or verb with its synonyms, hyponyms or hypernyms, double click on the word. Choose the desired WordNet relation, then select a node in the tree. The gloss for the selected synset and the expansion will show in the bottom panel. The gloss also is displayed as the tooltip for the node. The expansion consists of all the synsets beneath the selected node in the displayed tree. If you choose ‘use all word forms’ each word in the expansion will be expanded to all its morphological forms; otherwise, only the base form will be used.

3.3.9 Update Query and Cancel

The **Update Query** button updates the **Query** text box with the textual representation of the contents of the tree editor.

The **Cancel** button reverts the tree and the **Query** text box to their state when the page was most recently loaded.

3.4 Editing the Query

In addition to using the tree editor, you can manually edit the search in the **Query** text area. Clicking the **Update Tree** button will update the tree editor with your changes. If the parentheses in your query are not balanced you will receive an error message to this effect.

You can show and hide the query text box by clicking on the + or - icon next to it..

The **Update Tree** button will update the tree with the query contained in the

Query text box. If there is an error in the query (for example, unbalanced parentheses) this will be displayed instead of the query.

3.4.1 Morphological Expansion

To specify that a word should be expanded into all its morphological forms, add +VERB or +NOUN to the end. For example, `laughs+VERB` would be expanded to `laugh`, `laughs`, `laughing`, `laughed`. For WordNet expansions, +NOUN or +VERB should be specified after the synset but before the relation - for example `laugh#n+NOUN/hypo` or `cry#v#1+VERB/hype`.

3.4.2 WordNet Expansions

You can specify WordNet expansions directly in the query rather than using the tree editor. The full range of WordNet relations is available this way. We are currently using the WordNet 2.0 database. A web interface to WordNet is available at:

<http://www.cogsci.princeton.edu/cgi-bin/webwn>

Anything in the form `word#pos`, `word#pos#sense`, `word#pos/relation`, or `word#pos#sense/relation` will be expanded. If no relation is specified, the default relation is 'syns'.

POS is the part of speech and can be `n` for nouns or `v` for verbs.

Sense is the sense number in WordNet. For example, `body#n` has 9 senses in WordNet; `body#n#2` is "body, dead body - (body of a dead animal or person; 'they found the body in the lake')"

If no sense is specified, the relation will be applied to all senses.

An example:

`body#n#2/hpoi`

expands to

`mummy | carrions | cadaver | stiffs | stiff | mummies | remains | cadavers
| body | cremains | bodies | clay | carcass | carcase | remainises |
carrion | corpse | clays | carcasses | corpses | carcases`

The available relations are:

- `syns` - synset words
- `ants` - antonyms
- `hype` - hypernyms, brief (X is a kind of... / X is one way to...)

- **hypo** – hyponyms, brief (particular kinds of X) / troponyms (particular ways to X)
- **hpri** – hypernyms, full/inherited (all the things X is a kind of / way to, and all the things those are kinds of / ways to do)
- **hpoi** – hyponyms/troponyms, full/inherited (all the kinds of / ways to X, and all the kinds of / ways to do those things, etc)
- **mero** – all meronyms (regular) (parts of X)
- **mmem** – member meronyms (regular)
- **msub** – substance meronyms (regular)
- **mprt** – part meronyms (regular)
- **holo** – all holonyms (regular) (X is a part of...)
- **hmem** – member holonyms (regular)
- **hsub** – substance holonyms (regular)
- **hprt** – part holonyms (regular)
- **meri** – all meronyms, inherited (parts of X, and parts of those things, etc)
- **holi** – holynyms, inherited (things X is a part of, and things those things are parts of, etc)
- **attr** – attributes (...is a value of X)
- **enta** – entailment (verbs only)
- **caus** – cause (verbs only)
- **also** – also see
- **vgrp** – verb group (verbs only)
- **deri** – derived forms (nouns and verbs only)
- **domn** – domain – all
- **dmnc** – domain – category
- **dmnu** – domain – usage
- **dmnr** – domain – region
- **domt** – member of domain – all (nouns only)
- **dmtc** – member of domain – category (nouns only)
- **dmtu** – member of domain – usage (nouns only)
- **dmtr** – member of domain – region (nouns only)

3.4.3 Available Relations

The dominance and negation relations available in the tree edit can also be specified directly in the query.

Immediately Dominates Subtree This is the default relation. For example, the query (S NP) matches sentences with an S immediately dominating an NP.

Does not immediately dominate subtree To specify that a node should not immediately dominate the subtree, add a ! node. For example, the query (S (! NP)) matches sentences with an S that does not immediately dominate an NP.

Dominates subtree To specify that a node may appear in the sentence anywhere under its parent in the query, add a // node. For example, the query (S (// NP)) matches sentences with an S that has an NP as a descendant.

Does not dominate subtree To specify that a node must not appear anywhere in the sentence under its parent in the query, add a // node dominating a ! node. For example, the query (S (// (! NP))) matches sentences with an S that does not have an NP as a descendant.

3.5 Performing the Search

Once you are happy with your query, click **Update Query** to fill in the **Query** text box with the query you constructed in the tree editor. Then click the **Search** button. This will perform the query displayed in the **Query** text box with the options selected in the **Query Options** tab.

You can control several options about your search in the **Query Options** tab. To view this tab, click **Query Options**. You can select the source, the number of results to display per page, and control how the results are formatted.

3.5.1 Show Results As

Standard results simply display matching sentences along with a set of links for viewing the current version of the page on the web, an archived version of the page on the Internet Archive, and the annotations including parse tree for the page.

KWIC-formatted results allow you to view the results in a keyword-in-context style, centered on the word or phrase you enter. When using the KWIC result style you may need to scroll your browser horizontally to see all results.

3.5.2 Offensive Content Filter

If you are using or demonstrating the Linguist’s Search Engine for others or are otherwise in a setting where having the search engine return potentially offensive content could prove uncomfortable or embarrassing, you can turn on the Offensive Content Filter by checking the **Filter Offensive Content** checkbox. This applies a simple keyword-based scheme to sentences and URLs to hide results. If a particular result was filtered, its entry will still appear in the results but the sentence will be hidden and display “Filtered Content” instead. Unchecking the checkbox will immediately display any filtered results.

As with any keyword-based scheme, the filter is not foolproof, but it seems to work fairly well in practice.

3.5.3 Collection to Search

You can choose between one of the Public Collections (including the LSE Web collection) or one of “My Collections.”

The LSE Web Collection (public collection ‘lseweb’) is a corpus of about 3.5 million sentences chosen at random from Internet Archive crawls. This is the default source for English searches. There is also a much smaller Chinese Web collection (public collection ‘chinese_web’) of about 100,000 sentences of parallel Chinese/English text, also gathered from the Internet Archive. This collection is searchable in both English and Chinese and is the default collection for Chinese.

My Collections is the set of collections you have created using AltaVista searches (see 4). If you search My Collections, you must choose the particular collection to search on the main query page with the **Choose Collection** selection box. If you have not created any collections, only the LSE Web Collection and any Public Collections will be available to you.

Public Collections besides the LSE Web collection are similar to My Collections in that they are a narrower set of material, but are available to all users instead of just one user.

Currently available Public Collections include ‘bible,’ which is a parsed version of the World English Bible linked to dozens of other translations in English and many other languages. To view the other translations, perform a search, then select the ‘Annotations’ link from a result. You can choose which translations to see for Bible searches by following the ‘Select Translations’ link on the annotations page.

Once you have chosen a source, that will be the default for future queries until it is changed again.

3.5.4 Level to Search

Documents in some collections have been categorized automatically into "levels" corresponding to document difficulty, where 1 is the easiest level. The classification was done using a supervised classifier trained on manually "leveled" documents. Difficulty levels should be considered a noisy, experimental feature.

3.5.5 Downloading Query Results

You can download sentences returned by your query in two different formats. Both formats are a simple Comma-Separated Value (CSV) format that can be opened by any spreadsheet program. After clicking one of the **Download** buttons, you will be prompted to save the file. You should call it `results.csv` or something else with a `.csv` extension so that your spreadsheet program opens it properly.

Download Results simply gives you the URL to the document the sentence appears in, a link to the page on the Internet Archive and the body of the sentence.

Download KWIC-format results is a keyword in context format; the columns in the CSV file are the URL, a link to the page on the Internet Archive, the part of the sentence preceding the provided keyword, the keyword and the part of the sentence after the keyword. By opening the downloaded file, right-aligning the pre-keyword column and setting column widths appropriately you can view as much or as little of the sentences as you wish.

3.5.6 Query Results

An important note on the number of results displayed - when using the LSE Web Collection, the actual number of results *does not necessarily reflect* the true number of results that may be available. Querying the LSE Web Collection is a two-step process: first candidate results are generated and then the result set is refined. Currently, there is a hard limit of 1,000 candidates per search - candidates beyond this will be ignored. Not all of the information in your query is used when generating candidates, and the information used will differ from query to query; in some cases, there may be 1,000 or less candidates and all of them may be actual matches; in others, there may be many more than 1,000 candidates, but only a few of the first thousand candidates are actually matches.

For this reason, you *should not for any reason* use these result counts as an estimator of the frequency of your usage.

When using My Collections, the result count reflects the number of sentences out of those annotated that matched the query. Thus, estimation of frequency of usage *within the sentences retrieved by your AltaVista search* should be fairly accurate.

To navigate between pages of the result set, use the **Prev** and **Next** buttons.

There are three options for each result, each of which opens in a new browser window:

Annotation **Annotation** displays all available annotations for a sentence. This may include a constituency parse, dependency parse and part-of-speech tags. Not all annotations are available for all sentences.

From the annotation page, you can select a parse to be loaded into the tree editor by clicking the “Use This Sentence for Query By Example” link.

Archived **Archived** retrieves the page using the Internet Archive. If the page is a part of the LSE Web Collection, it will attempt to retrieve the version of the page used in the LSE Web Collection. This version may not always be available.

If the page was part of a collection created with My Collections, all versions of the page the Internet Archive knows about will be displayed.

Current **Current** loads the page as it is currently on the Web. The page may have been changed or been removed since it was downloaded by the Linguist’s Search Engine.

3.6 Saving and Loading Queries

The query interface allows you to save queries for future use - perhaps you originally use a query on the LSE Web Collection and want to build a collection likely to give more results using My Collections. You can save the query for later use and apply it to your new collection. Or, you may want to bookmark an interesting result set.

3.6.1 Saving Queries

To save a query, make sure the query you want to save is displayed in the **Query** text box. Click **Save Query** to display the Save Query tab. Enter a display

name and a short description for the query in the **Name** and **Description** text boxes and click **Save Query**. A message will be displayed informing you that your query has been saved.

3.6.2 Using Saved Queries

To use a saved query, first click **Load Query** to display the **Load Query** tab. You can sort your saved queries by name or by modification date. To display the description for a saved query, click on the + icon next to a particular query. To load a saved query, click on the name. The tree editor and the query text box will be updated with your query. Loading a saved query will not affect the search options. To delete a saved query, click the + icon next to the particular query and then click the **Delete** button.

To change a saved query, first load the saved query, then click **Save Query** to display the **Save Query** tab. The name and description for your query will be shown. If you change the query or the description but keep the name, saving the query will overwrite the old saved query with that name.

4 Collection Building

The Linguist's Search Engine can automatically build a searchable collection of sentences based on the results of an AltaVista search.

4.1 Creating a Collection

To create a new collection, select **My Collections** from the navigation bar and click **Add New Collection Definition**.

You must enter a name for your collection in the **Collection** textbox as well as a description in the **Description** text box.

To add results from a new AltaVista search to your collection, enter the search to be performed on AltaVista in the **AltaVista Search Terms** text box. The syntax for this search is the AltaVista Advanced Search syntax. In addition you can use **word+VERB** and **word+NOUN**, which will expand to all verb or noun morphological forms of **word** respectively. WordNet relationships (see 3.3.8) are also supported here.

Most likely, you will not be interested in all the sentences in all the documents retrieved by your search - you only want sentences that contain some particular word. Enter a space-separated list of words in the **I only want sentence that contain..** textbox - only sentences that contain one or more of the words will be added to your new collection. You can also select the maximum number of

documents to retrieve from the list of results returned by AltaVista. You can also use **+VERB**, **+NOUN** and WordNet relationships in this field - these will be shown in their expanded form once you save changes.

Before saving changes and adding your new search to your collection, you should click **Try This Query** to view the results of your search. This will alert you to any problems with your search. You can also use the **Try this query on AltaVista** link for searches you have already created.

Once you are happy with your search, click **Save Changes** to save your search or **Start Annotating** to start downloading and annotating sentences.

4.2 Managing a Collection

4.2.1 Search Display

Once you have added a search, you will see the time the search was created, the actual AltaVista search, and the must-contain-one-of word list. If the search hasn't been performed yet, you can select the **Delete this query** checkbox and click the **Save Changes** button to remove the query. Any **+VERB**, **+NOUN** or WordNet relationships will be shown in their expanded form here. If you are not happy with the expansion you can delete the search and create a new one with the correct expansion.

4.2.2 Annotation Priority

The annotation of your collection is a computationally intensive process that competes for resources with other users of the Linguist's Search Engine. To ensure a fair use of resources, each AltaVista search job is assigned a priority. This priority increases as the search gets older and decreases as sentences are annotated. This means (for example) that a new search with no sentences annotated would have a higher priority than a day-old search that already has 10,000 sentences annotated.

4.2.3 Status

There are several possible statuses for your collection. In addition to these statuses, you will be informed as to the number of sentences found and the number of sentences annotated.

Awaiting 'Start Annotating' command You have created a search but have not yet clicked **Start Annotating**. You should make sure all the searches

you have created return appropriate results from AltaVista before clicking **Start Annotating**.

Queued The annotation process is started, but your search has not yet been run.

Downloading Your search has been run on AltaVista and result pages are being downloaded.

Processing Result pages are being processed - each document must have HTML tokens removed, sentence boundaries detected, etc. The must-contain-one-of word list is used in this step to find sentences of interest.

Annotating Sentences of interest have been found and are being parsed.

Finished All the sentences found have been parsed.

4.2.4 Stopping Annotation

You can search your collection as soon as there are sentences annotated - if you find what you're interested in as soon as 500 sentences are parsed and there are 20,000 more sentences to annotate, you can use the **Stop Annotating** button to halt the annotation process.

4.2.5 Delete This Collection

If a collection is no longer of interest, click **Delete This Collection** to completely remove the collection. This can take several minutes depending on the size of your collection.

4.3 Searching Your Collection

You can search your collection as soon as there are sentences annotated - there is no need to wait for all the sentences to be annotated, though of course only the sentences that have finished being annotated will be searched. To search your collection, select **My Collections** as the source on the Query by Example or Query pages, then select the appropriate collection under **Collections**.

If you already know the query you want to run, you can also use the **Search This Collection** button on the collection detail or collection list pages.

4.4 Listing Your Collections

A summary view of your collections is displayed on the main **My Collections** page. This page shows the collection name, date of creation and brief status for each of your collections. You can sort by date and name by clicking on the column headings. Use the + icon to expand the view for a particular collection and see the description, full status and details of each AltaVista search. Use the **Edit** button to edit the searches for each collection, add new searches or to delete the collection.