Linguistic Search Engine
Searching Tagged Hebrew Text

Nathan Grunzweig and Yoav Goldberg and Michael Elhadad

Ben Gurion University of the Negev

ISCOL 2010

Corpus Linguistics

- Data driven methodology:
  - Collect dataset
  - Manually inspect data
  - Tag the data
  - Learn from the data
  - Analyze errors
- Searching tagged corpora is useful but difficult

Tagged Corpus

The/AT **grand**/JJ **jury**/NN **commented**/VBD
on/IN **a**/AT **number**/NN **of**/IN **other**/AP **topics**/NNS **,**/,
AMONG/IN **them**/PPO
the/AT **Atlanta**/NP **and**/CC **Fulton**/NP-tl **County**/NN-tl **purchasing**/
VBG **departments**/NNS
which/WDT **it**/PPS **said**/VBD **``**/`` **ARE**/BER **well**/QL **operated**/VBN
and/CC **follow**/VB **generally**/RB **accepted**/VBN **practices**/NNS **which**/
WDT **inure**/VB **to**/IN **the**/AT **best**/JJT **interest**/NN **of**/IN **both**/ABX
governments/NNS **"**/" **.**/.

Interesting queries:

Sentences containing W tagged as T
2 proper nouns followed by a definite feminine adjective
A past verb followed by a masculine noun, with no more than two words between them

Searching Corpora

Standard search engines are not directly adapted to tagged corpora:

- Document vs. sentence granularity
- Support mixed queries tag/words
- Avoid intrusive indexing methods (stemming, tokenization, query expansion)
- Control content of the corpus

Our Linguistic Search Engine

Based on the open source Lucene platform

Added the capability of indexing based on both words and their many morphological properties.

Efficient search over orthographic word forms, as well as linguistic properties (part-of-speech, lemma, gender, tense)

Hebrew Support

- Text is automatically tagged using the BGU morphological disambiguator and chunker

- We indexed about 110M tokens (~8M sentences) from various genres including blogs, news, Knesset proceedings and medical articles.

Query Language

We designed a specific query language that combines words and their properties

- Word is X and gender is Y
  - $w.form="x" & $w.gender="y" ; $w

- Word is X and gender is Y inside Z
  - $w.form="x" & $w.gender="y" ; [$w # "z"]

- Word X 3 words after word Y
  - $w.form="x" & $w1.form="Y" ; $w !~3 $w1

Query Compilation

Queries are compiled into a Lucene query, for example:

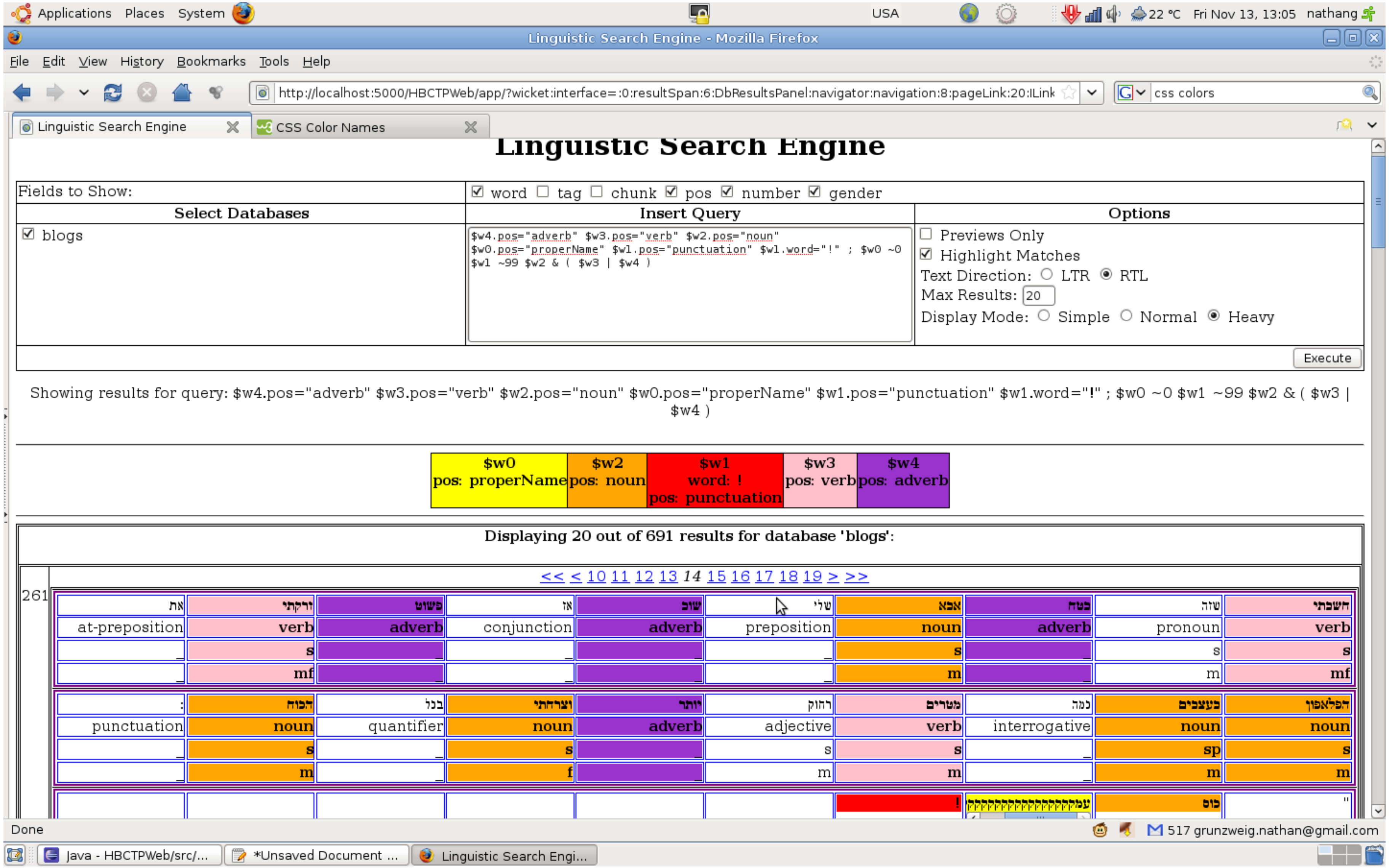
- $w.word="אני" & $w1.pos=“adjective” & $w2.word=“נתן” & $w2.pos=“properName” & prefix=“ה*” & suffix=“ל*";
  [$w ~0 & ($w1 | $w2) # “NP” + $prefix + $suffix] ~0 pos=“verb”

spanNear([spanNear([spanWithin(spanNot(spanOr([spanNear ([property:word= א, property:pos=verb], 0, false), spanNear ([property:word= א, spanNear([property:word= נ, property:pos=properName], -1, false)], 0, false)]), spanOr([property:@E-N-D@=NP, property:@S-T-A-R-T@=NP])), 1 ,spanNear([spanNot(spanNear ([property:@S-T-A-R-T@=NP, spanWildcardQuery(property:word= ה, true), property:@E-N-D@=NP), spanNot(spanNear([spanWildcardQuery (property:word=ל*), property:@E-N-D@=NP], 0, true), property:@S-T-A-R-T@=NP)], 9999, true)), spanNear([spanNot(spanNear([property:@S-T-A-R-T@=NP, spanWildcardQuery(property:word= ה, true), property:@E-N-D@=NP), spanOr([spanNear([property:word= א, property:pos=verb], 0, false),

……

Additional Features

- Web Application (embedded in Jetty server)
- Match highlighting (multicolored)
- Highlighting Legend
- Several result formats
- RTL\LTR
- Selectable properties to display



Applications Places System
USA
22 °C Fri Nov 13, 13:05 nathang
Linguistic Search Engine - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://localhost:5000/HBCTPWeb/app/?wicket:interface=:0:resultSpan:6:DbResultsPanel:navigator:navigation:8:pageLink:20:ILink
css colors
Linguistic Search Engine
CSS Color Names

Linguistic Search Engine

Fields to Show: ☑ word □ tag □ chunk ☑ pos ☑ number ☑ gender

Select Databases	Insert Query	Options
☑ blogs	`$w4.pos="adverb" $w3.pos="verb" $w2.pos="noun" $w0.pos="properName" $w1.pos="punctuation" $w1.word="!" ; $w0 ~0 $w1 ~99 $w2 & ( $w3 | $w4 )`	□ Previews Only ☑ Highlight Matches Text Direction: ○ LTR ● RTL Max Results: 20 Display Mode: ○ Simple ○ Normal ● Heavy

Showing results for query: $w4.pos="adverb" $w3.pos="verb" $w2.pos="noun" $w0.pos="properName" $w1.pos="punctuation" $w1.word="!" ; $w0 ~0 $w1 ~99 $w2 & ($w3 | $w4)

$w0 pos: properName	$w2 pos: noun	$w1 word: ! pos: punctuation	$w3 pos: verb	$w4 pos: adverb

Displaying 20 out of 691 results for database 'blogs':

<< < 10 11 12 13 *14* 15 16 17 18 19 > >>

261

את	זרקתי	פשוט	אז	שוב	שלי	אבא	בטח	שזה	חשבתי
at-preposition	verb	adverb	conjunction	adverb	preposition	noun	adverb	pronoun	verb
_	s	_	_	_	_	s	_	s	s
_	mf	_	_	_	_	m	_	m	mf

:	הכוח	בכל	וצרחתי	יותר	רחוק	מטרים	כמה	כעצבים	הפלאפון
punctuation	noun	quantifier	noun	adverb	adjective	verb	interrogative	noun	noun
_	s	_	s	_	s	s	_	sp	s
_	m	_	f	_	m	m	_	m	m

						!	עמקקקקקקקקקקקקקקקקק	כוס	"

Availability

- Indexed corpora are available at:

  http://www.cs.bgu.ac.il/~yoavg/nathan/HBCTPWeb/app

- Software available