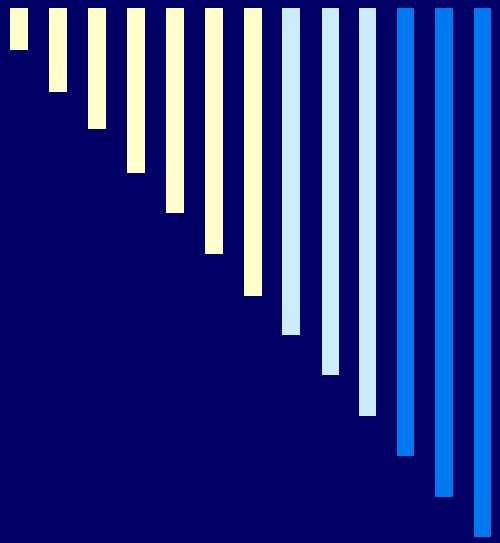
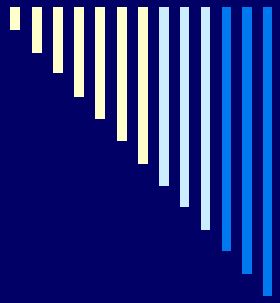


A computational Lexicon for Contemporary Hebrew

Alon Itai – CS Technion

Shuly Wintner – CS Haifa University

Shlomo Yona – CS Haifa University



Outlook

- Modern Hebrew

- What is a lexicon?
- What is in our lexicon?
- Why do we need it?
- How did we acquire it?

Modern Hebrew

- Official Language of the State of Israel
- Spoken by 7 M people
- Related, but linguistically distinct, from Biblical Hebrew.

Semitic Word Formation

root + pattern → word

pattern / root	CaCaC	yiCCoC	hitCaCCeC
ktb	katab (he wrote)	yiktob (he will write)	hitkatteb (corresponded)
šbr	šabar (he broke)	yišbor (he will break)	hištabber (refract)

Writing System

- Most vowels are omitted
- Particles are prepended to words,

Example:

h – definite article,

b – preposition (in)

w – conjunction (and)

wbbyt = w + b + ha +byt

and in the house

Morphological Ambiguity

- Most words are morphologically ambiguous
- Example: šbth שבתה

1. šavta = šbt + CaCCa = stopped working
2. šavta = šbh + CaCCa = took prisoner
3. šabatah = her Saturday
4. še-b-te = that in tea
5. še-b-ha-te = that in the tea
6. še-bit-h = that her daughter

…

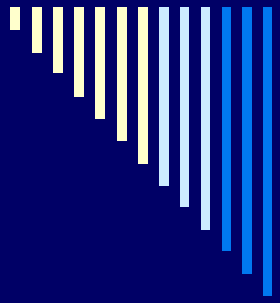
How to morphologically parse?

- Create all patterns
- Given a token – check whether it fits a pattern.

Example: In English xxxs → xxx (noun) + s
houses → house; *bosses → bosse

- Creates a lot of superfluous parses.
- Use a lexicon to reduce the number of parses

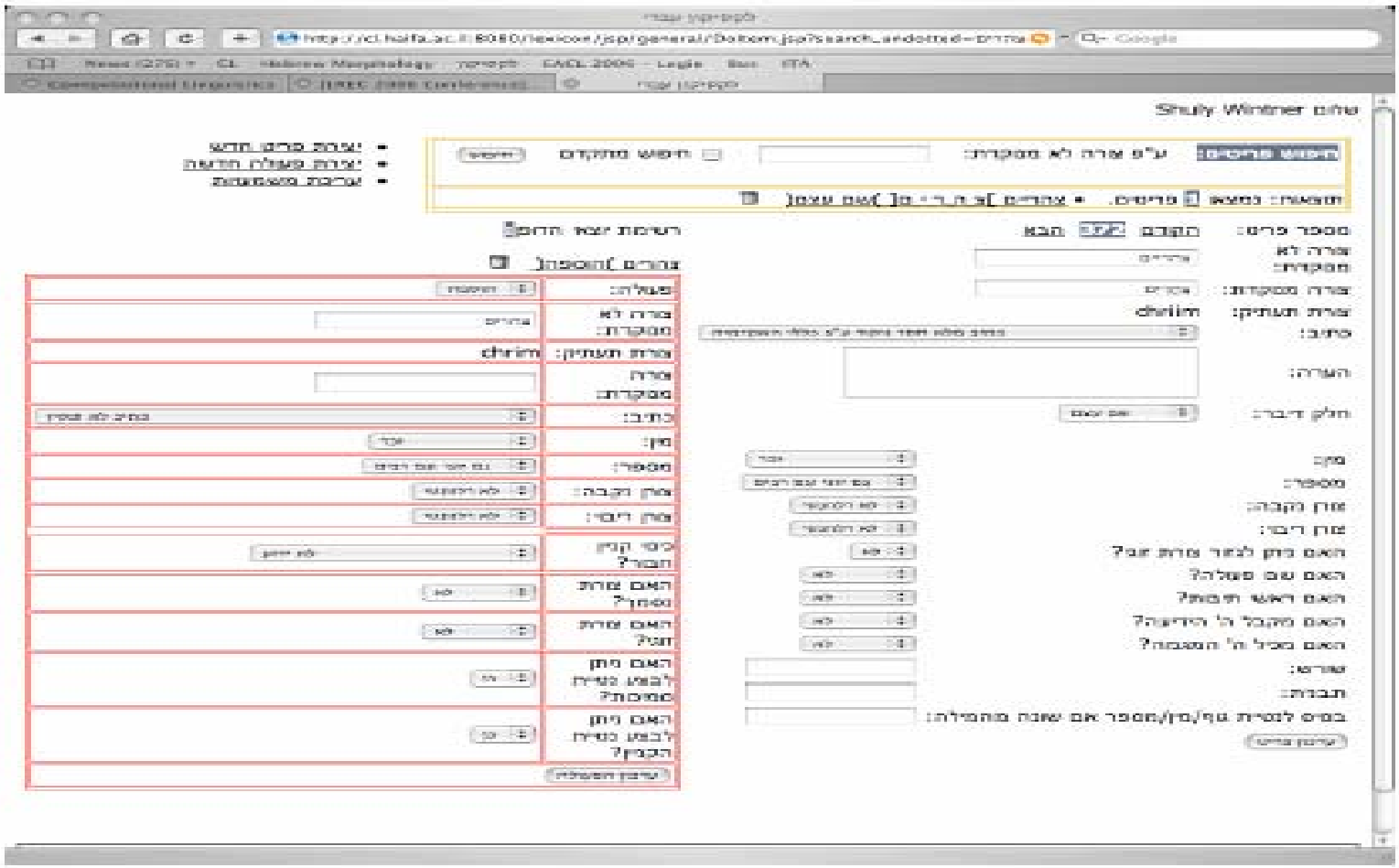
bosse $\notin$ lexicon

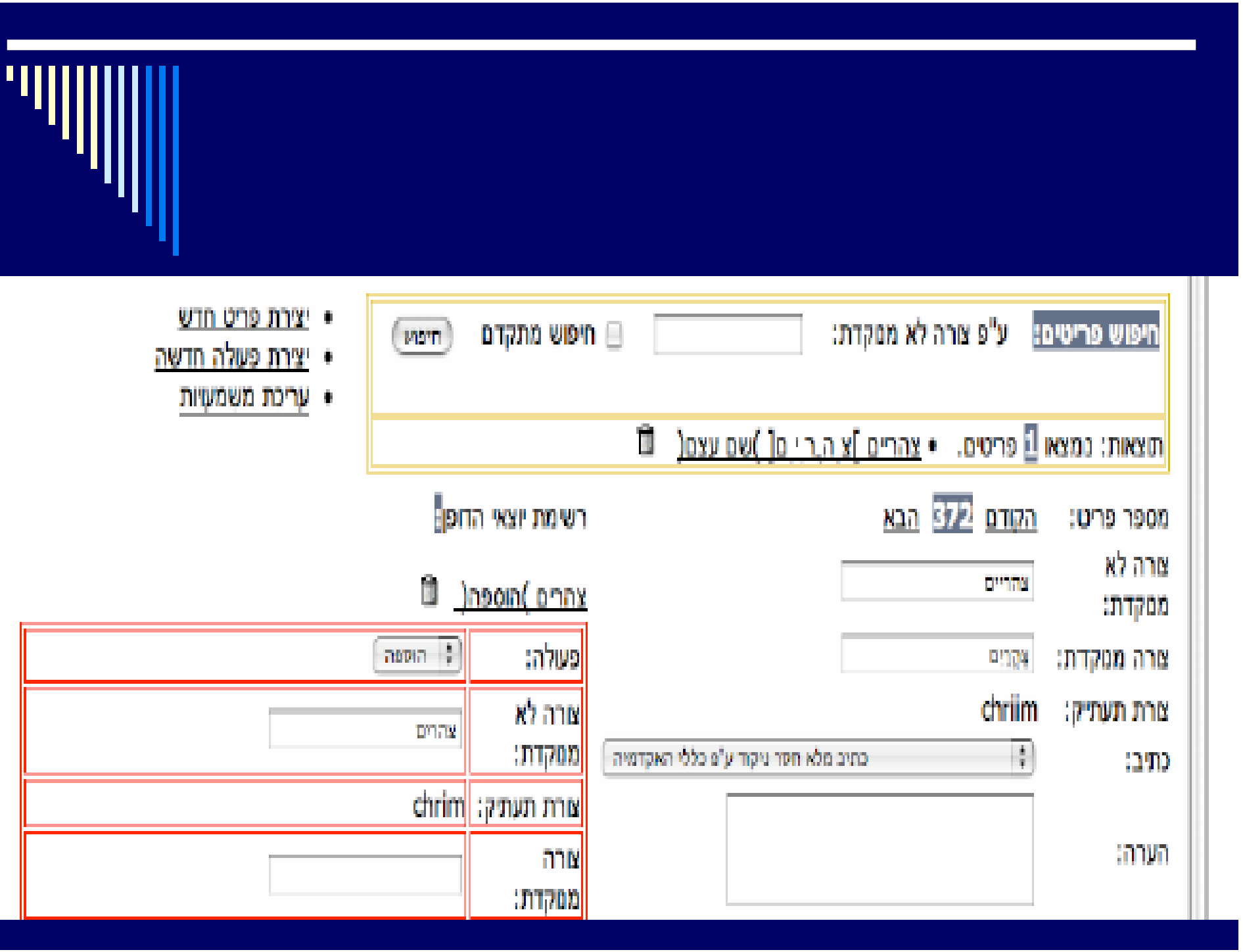


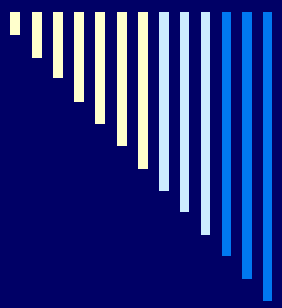
Acquisition

- Started with lexicons of previous morphological analyzers (HSPELL, Segal).
- Added missing conjugations, such as passives, and nomalizations (manually verified).
- Parsed corpora and listed tokens that had no morphologically valid parse. (Mainly proper names). Added them (manually to the lexicon).

GUI for editing the lexicon




חיפוש פריטים:
ע"פ צורה לא מסקדת:
חיפוש מתקדם
חיפוש
תוצאות: נמצאו 1 פריטים.
צהריים]צ ה ר י ם[ )שם עצם(
יצירת פריט חדש
יצירת פעולה חדשה
עריכת משמעויות
מספר פריט:
הקודם 372 הבא
רשימת יוצאי החפץ
צורה לא מסקדת:
צהריים
צהרים)הוספה(
צורה מסקדת:
צהרים
פעולה:
הוספה
צורת תעתיק:
chriim
צורה לא מסקדת:
צהרים
כתיב:
כתיב מלא חסר ניקוד ע"פ כללי האקדמיה
צורת תעתיק:
chriim
הערה:
צורה מסקדת:



Size of the lexicon by part of speech

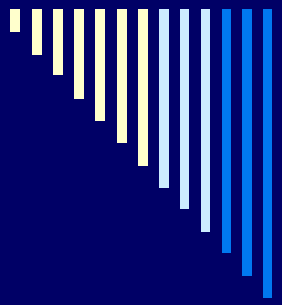
noun	10332	preposition	100
verb	4485	conjunction	62
Proper Name	4227	pronoun	60
adjective	1612	interjection	40
adverb	352	interrogative	9
quantifier	132	negation	6

Total : 21,417

Organization

- Ordered by lexeme, not root.
- Similar to nearly all dictionaries.
- Most laymen cannot identify the root.
- The semantics is associated with the lexeme and only loosely with the root

- paqad – visited hitpaqqed
  nifqad – missing hifqid -deposited
  piqqed -- commanded



Structure of an entry

- Unique ID

Nominals: (nouns, adjectives)

- The lexical item: dotted, undotted, transliterated
- POS
- Gender / number
- Plural suffix (im, ot).
- Inflection base (if different)
- Exceptions (if inflection has exceptions)

Structure of an entry (2)

Verbs

- Root
- Inflection pattern = binyan + pattern of 1st binyan
  škb + tiCCC → tiškb (tiškav)
  psl + tiCCC → tipsol (tifsol)
- Valency

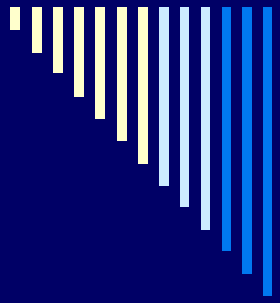
XML

- The lexicon is represented in XML
- Readable both by machines and by humans
- Enables using off-shelf tools for on screen presentation and validation

EXAMPLE

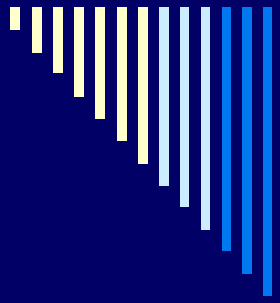
```
-<item id="17580" script="formal" transliterated="bwqr"
   undotted="בוקר" dotted="בֹּקֶר" >
  <noun gender="masculine" number="singular" plural="im">
    <replace gender="masculine" number="plural" script="formal"
      transliterated="bqarim" undotted="בקרים"/>
  </noun>
</item>
```

Info for the morphological parser



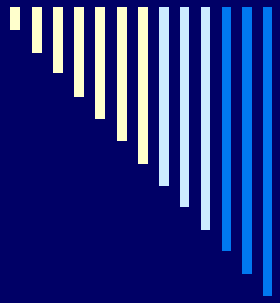
License

- Available under GPL – Gnu Public License. You get it for free if all products derived from it are also under GPL.
- Can get a non-exclusive license for commercial use.



Conclusions

- Created a comprehensive lexicon of Modern Hebrew.
- Identify 96% of all tokens in corpus.
- Missing: Proper names, typos, nonstandard spelling, …
- Open for research under GPL
- Created within the Knowledge Center for Processing Hebrew



Acknowlodgements

- Knowledge Center for Processing Hebrew
- Israel Ministry for Science and Technology
- People:
  Shuly Wintner – Haifa University
  Shlomo Yona – Haifa University
  Yoad Winter – Technion
  Shira Schwartz – lexicographer
  Dalia Bojan – software engineer