

# Electronic Corpora for Two Semitic Languages

---

Jerid Francom, Wake Forest University, Winston-Salem, USA  
Dainon Woudstra/ Adam Ussishkin, University of Arizona, Tucson, USA

October 10, 2009

AAACL 2009

University of Alberta, Edmonton

<http://psycol.arizona.edu>

# Acknowledgements

---

- Generous contributions to this project by
  - MILA Knowledge Center for Processing Hebrew
  - Dr. David Plaut (Carnegie-Melon University)
  - Dr. Albert Gatt (U of Aberdeen/ U of Malta)
- Funding from the United States National Science Foundation (BCS-0715500) to Adam Ussishkin

# Overview

---

- Project
  - Creation of a Maltese & Hebrew lexical corpus and web-interface to these collections
  - Why Semitic? Our lab is interested in language processing for languages with nonconcatenative morphology.
- Goal
  - Develop sizeable corpora and a set of tools capable of providing lexical statistics relevant for psycholinguistic research

# Hebrew corpus

---

- Data acquisition
  - Newspapers articles & opinion, TV transcripts & medical forums
    - *2b-bari, Arutz7, Ha'aretz, Doctors, Infomed Medical Forum, Tapuz People Forums, The Marker & Ynet*
  - Date range: 2001-2006
  - MILA Knowledge Center for Processing Hebrew & David Plaut
- Data indexing
  - Tokenized and added to SQL database

# Hebrew corpus

---

	Tokens	% of collection
Sources		
2b	709,024	1.18%
A7 articles	1,220,090	2.03%
Doctors	196,603	0.33%
Ha'aretz	8,273,572	13.78%
Infomed	163,649	0.27%
Tapuz	1,004,998	1.67%
TheMarker	559,438	0.93%
Ynet	47,924,887	79.81%
Totals		
	60,052,261	Type/token ratio 0.7%

# Maltese Corpus

---

- Data acquisition
  - Web crawled a set of Maltese newspapers
    - Illum
    - L-Orizzont
    - Malta Right Now
  - Inconsistent rendering of Maltese characters  $\dot{c}$ ,  $\dot{g}$ ,  $\dot{h}$ , and  $\dot{z}$ , forced us to eliminate some possible sources.
  - Date range: 2005 - 2007
  - Employed *Wget* to retrieve the web data
- Data filtering & indexing
  - Filtered, tokenized and added to SQL database

# Maltese Corpus

---

- Supplementary data
  - Newspaper data acquired from Albert Gatt
    - Kullhadd
    - In-Nazzjon
    - Lehen is-Sewwa
  - Date range: 1998 - 1999
- Indexing
  - Tokenized and added to SQL database

# Maltese corpus

---

	Tokens	% of collection
PsyCoL Web crawl		
Illum	1,927,598	58%
L-orizzont	60,982	1.8%
Provided by A. Gatt		
Kullhadd	69,908	1.8%
In-Nazzjon	1,240,923	37.3%
Lehen is-Sewwa	23,914	0.7%
Totals		
Web crawl	1,988,580	59.8%
A. Gatt	1,334,745	40.2%
	3,323,325	Type/token ratio 1.6%



# Description of the corpus interface

---

- Goals
  - Provide international, cross-platform access to the lexical corpora
- Tools
  - General workbench
  - Specific information extraction tools
    - Lexical frequency
    - Uniqueness point
    - Neighborhood density

# General tools

- Documentation  
<http://psycol.arizona.edu>
- Language selector
- Search field  
*(RegExp enabled - important for Semitic)*
- Virtual keyboard

## PsyCoL Maltese Lexical Corpus (PMLC)

### Sources

The PsyCoL Maltese Lexical Corpus (PMLC) is composed of on-line newspapers within two general data ranges: 1) 1998 - 1999 and 2) 2005 - 2007.

All data was retrieved from the web but the two data ranges cited here reflect two distinct efforts. The first by Albert Gatt who has graciously shared data he collected from various sources (Kulhadd, Lehen, Il-Mument and In-Nazzjon). This work represents 1,395,727 tokens and 53,396 unique types. The second effort was conducted by the PsyCoL lab and includes the all other data collected (Illum, <Malta Right Now>) which adds 1,927,598 <+> tokens to the corpus.

All data was converted and tokenized in UTF-8 and is stored in a relative database structure (MySQL v. 5.0). Each data source below is represented as a single column in the table structure with values that correspond to the full set of unique tokens found across all sources aggregated here. In addition, there is a column for the total token count for each unique token.

[View Source Details](#)

The screenshot shows the PsyCoL website interface. At the top, the logo "PsyCoL" is displayed in red and green, with the text "Psycholinguistics & Computational Linguistics Lab" below it. On the right, there is a user greeting "Hello Jerid!" and links for "Settings" and "Logout". A navigation bar contains tabs for "HOME", "PROJECTS", "LANGUAGE RESOURCES", "TOOLS", "PERSONNEL", "FACILITIES", and "SITE SEARCH". Below the navigation bar, there are links for "Token Frequency Calculator", "Neighborhood Density Calculator", and "Lexical Uniqueness Point Calculator". The "Token Frequency Calculator" tool is active, showing a dropdown menu for language selection with "Maltese" selected. Below the dropdown, the text "Token Frequency Calculator" is displayed, followed by "Total Token Count: 3,323,325 | Unique Tokens: 53,396". A virtual keyboard is visible below the text, with keys for letters, numbers, and symbols. At the bottom, there is a "Calculate" button, a text input field containing "kumpaniji", and a "Reset" button.

# Specific tool: lexical frequency

- What it measures
- Use in research  
Frequency plays a role in lexical access, as shown by numerous studies.

In general, the more frequent a word, the easier it is to retrieve.

Results for: **tuffieħa**

Index	Token	Count Per Million	Natural Log	Db Count	Kullhadd	InNazżjon	MaltaRightNow	Lorizont	Lehen_isSewwa	%	Query Total
17668	tuffieħa	1.805	1.79176	6	0	6	0	0	0	2E-06	6

Query Specific Counts

Db Unique Tokens	1
Query Specific Tokens	6
<i>Kullhadd</i>	0
<i>InNazżjon</i>	6
<i>MaltaRightNow</i>	0
<i>Lorizont</i>	0
<i>Lehen_isSewwa</i>	0

# Specific tool: uniqueness point

---

- What it measures
- Use in research

The lexical uniqueness point of a given word plays a role in lexical access (e.g., Marslen-Wilson 1978, Wurm 2007).

Lexical access of an auditorily-presented word can proceed more deterministically from the point in time corresponding to the lexical uniqueness point of the word.

**Results for: tuffieħa**

Word	Left Index	Right Index
tuffieħa	8	1

**Results for: כמה**

**Word: Not Unique**  
36 Overlapped words  
Word list: כמהמכרות, כמהמר, כמהמהם, כמהאבות, כמהנתונים, כמהנדס, כמהנדסים, כמהנדרשת, כמהסת, כמהפאזלים, כמהה, כמהפנט, כמהפכו, כמהפכניות, כמהפכה, כמהתלה, כמהתלים, כמהתוכניות, כמהלך, כמהלומה, כמהגרים, כמהדורה, כמהה, כמהההההה, כמהו, כמהופנט, כמהופנטים, כמהות, כמהותה, כמהותיים, כמהויים, כמהין, כמהימן, כמהימנה, כמהימנים, כמהים, כמהיר

# Specific tool: neighborhood density

---

- What it measures
- Use in research

Neighborhood density also plays an important role in lexical access (e.g., Goldinger, Luce, and Pisoni, 1989; Cluff and Luce, 1990; Luce and Pisoni, 1998).

In visual studies, higher neighborhood density correlates with faster lexical access.

In auditory studies, higher neighborhood density correlates with slower lexical access.

## Results for: tuffieña

<b>Density Measures:</b>	
Number of neighbors: 1	
Frequency of this neighborhood in corpus: 9	
Neighborhood count per million: 2.7081312	
Natural log frequency: 2.1972246	
<hr/>	
<b>Neighborhood Statistics</b>	
<b>Word</b>	<b>Neighbors</b>
tuffieñ	3
<b>Query Word</b>	<b>Frequency</b>
tuffieña	6

# Project assessment

---

- Size

PsyCoL Hebrew Lexical Corpus (PHLC)

Total Token Count: 60,052,261 Unique Tokens: 396,469

PsyCoL Maltese Lexical Corpus (PMLC)

Total Token Count: 3,323,325 Unique Tokens: 53,396

- Accessibility

Web interface

- Extensibility

Khalkha Mongolian already added

Additional tools can be integrated

# Conclusion

---

- Creation of lexical corpora for Maltese & Hebrew
- Developed of a set of tools particularly useful for psycholinguistic research
- Contributes to growing body of research to connections between corpus and behavioral inquiry.

Thank you!