

## Lexical access in Maltese and Hebrew: The intersection of corpus linguistics and psycholinguistics

Adam Ussishkin

University of Arizona, Department of Linguistics  
PsyCoL Lab



# Acknowledgments

- Many thanks to the United States National Science Foundation for supporting portions of this research (NSF award BCS-0715500).
- I also offer thanks to:
  - Jeff Berry (Univ. of Arizona)
  - Dr. Ray Fabri (Univ. of Malta)
  - Dr. Jerid Francom (Wake Forest Univ.)
  - Dr. Ram Frost (Hebrew University of Jerusalem)
  - Dr. Albert Gatt (Univ. of Malta)
  - Amy LaCross (Univ. of Arizona)
  - Hadas Velan (Hebrew University of Jerusalem)
  - Dr. Andy Wedel (Univ. of Arizona)
  - Dainon Woudstra (Univ. of Arizona)

# The question

- How does lexical access work in Semitic languages?
- Rather than answering this question definitively, today I'll be describing research components involved in approaching the question.
- These components include details on corpus creation, corpus testing, and behavioral studies.

# Experiment 1: Family size and frequency effects

- We tested lexical access in Modern Hebrew using auditory stimuli to test effects of 3 psycholinguistic variables:
  1. Word frequency (based on the David Plaut printed Hebrew corpus, 200 million tokens)
  2. Related family size (RFS) and
  3. Unrelated family size (UFS)
- Auditory replication of Experiment 1 in Moscoso del Prado-Martín et al. 2005 (henceforth referred to as “MPM”) to avoid orthographic confound.

# Experiment 1: Hypotheses

- A strictly root-based theory predicts no word frequency effects.
- A strictly word-based theory predicts no family size effects for consonantal root families.
- A theory that recognizes both roots and words predicts that both variables would have effects.

# Related vs. unrelated family size

- The two Hebrew words [ʃémɛn] ‘oil’ and [ʃamén] ‘fat’ share a consonantal root, and are semantically closely related.
- In contrast, the two Hebrew words [xɑʃáv] ‘he thought’ and [xɛʃbón] ‘account’ share a consonantal root, but are much more distant semantically.
- MPM used Latent Semantic Analysis (Landauer and Dumais 1997) to determine and quantify degree of semantic relatedness.

# Auditory lexical decision

- Our experiment manipulated these same variables (RFS, UFS, and word frequency) using materials identical to those used by MPM.
- Rather than displaying printed stimuli on a computer screen, our subjects responded to stimuli that were presented auditorily over headphones.

# Materials

- Materials consisted of 99 real words and 99 nonwords (nouns and verbs), identical to the materials used by MPM.
- For real words:
  - ▣ 43 from non-homonymic roots – i.e., having only related family members
  - ▣ 56 from homonymic roots – i.e., having both related and unrelated family members
- All nonwords were created from *nonce roots* – i.e., unattested but phonologically legal consonantal roots in Hebrew.
- All real word stimuli differed with respect to three variables: word frequency, RFS, and UFS.

# Participants, procedures, and apparatus

- All native speakers of Hebrew with normal hearing
- Students at the Hebrew University of Jerusalem; experiment run in the Verbal Processing Laboratory
- 118 participants received course credit or monetary compensation
- Stimuli recorded by a female native speaker of Hebrew, using a head-mounted microphone in a sound booth in the Douglass Phonetics Lab at the University of Arizona
- Experiment conducted on IBM PC, using DMDX software (Forster and Forster 2003) to measure RT and accuracy in an auditory lexical decision task
- Participants wore headphones to hear stimuli and pressed one of two response buttons on a keyboard “YES” or “NO”

# Results

- LMER analysis
- Significant negative effect of word frequency (coeff=14.37 ms per log unit)
  - ▣ Higher word frequency facilitates lexical access.
- Significant positive effect of UFS (coeff=6.1 ms)
  - ▣ Higher UFS inhibits lexical access.
- No effect in LMER of RFS (but a significant effect of RFS does turn up in a multiple regression analysis)

# Experiment 1: discussion

- Family size effect implicates representation of consonantal roots.
- Word frequency effect implicates representation of words.
- Obviously, this isn't the end of the story, but these results are highly suggestive of a model in which both roots and words are lexically represented.

# Midpoint meta-discussion

- Big issue: lexical organization and structure in Semitic.
- Important confound: Most Semitic languages are written with a consonantal orthography.
- Exp 1 attempts to mitigate the confound using auditory instead of visual stimuli, but the confound isn't altogether eliminated.
  - Orthography can influence the structure of the lexicon (Forster 1976, Morais 1986), so auditory experiments in Hebrew may not succeed in avoiding the confound.

# Midpoint meta-discussion

- Ideal: a Semitic language with an orthography that doesn't implicitly favor consonants: Maltese
- Problem: Maltese was seriously under-resourced.
  - No comprehensive online dictionary.
  - No word frequency or familiarity data.
  - No neighborhood density data.
  - No lexical uniqueness point data.
- Solution: build a corpus, and dive in.

# The PsyCoL corpora

- We created two Semitic corpora: one for Maltese, and another for Hebrew.
- Modern computational tools make corpus creation straightforward.
  - ▣ The web provides a natural and accessible source for a corpus.
  - ▣ From a practical standpoint, the web contains a huge amount of text in electronic form, thus obviating the need for time-consuming pre-processing and tedious conversion of print-to-electronic text.

# Web as corpus

- The web also presents potential pitfalls.
  - ▣ Lack of representativeness: the text on the web may not be a perfect sampling and may in fact be limited to particular domains.
  - ▣ Not all languages are equally represented: 90% of web-based text comes from a small handful of languages (Xu 2000, Kilgarriffe and Grefenstette 2003).
  - ▣ High variance: lack of rigorous publishing standards for web-based text and special character rendition present problems.

# Web as corpus

- Nonetheless, the web represents a level of accessibility and scalability that may overcome these potential pitfalls.
- In what follows, I'll detail the development of our corpora through the necessary stages.

# Seed selection

- This first step involves selecting the online sources for our corpora.
- A pre-selection list was created after a Google search.
  - This yielded newspapers and blogs as the most data-heavy and accessible sources of Maltese and Hebrew text.

# Seed selection

- Several potential problems helped us decide on which sources to use.
  - ▣ Amount of text in potential source
  - ▣ Non-target language contamination
  - ▣ Proofreading standards
- As a result, our corpora are mainly based on newspaper sources.

# Seed selection

- Maltese:
  - Character encoding revealed itself to be a decisive factor, and forced us to eliminate some possible sources.
  - For instance, the Maltese characters ċ, ġ, ħ, and ż are not consistently rendered on some newspaper websites.
  - Remaining candidates included Illum, L-Orizzont, Malta Right Now.
  - Data were also generously provided by Dr. Albert Gatt from Kulĥadd, Leĥen is-Sewwa, and In-Nazzjon.
- Hebrew:
  - Instead of re-inventing the wheel, we were provided raw text from the Mila Center for Processing Hebrew as well as from Dr. David Plaut.

# Web extraction (Maltese)

- Once seed selection for the corpus was complete, we proceeded to extraction from the web (for Maltese).
- We used the `Wget` command in UNIX to crawl our seed sites.
- Text between certain types of html tags was eliminated in an effort to filter out non-target language text.

# Tokenization

- All data were split using text-based markers that reflect morphological/word boundaries, including the apostrophe and non-alphabetic characters (hyphen, slash, and a few others).
- The result is a word list that is then tokenized.
- The resulting array was then re-processed into a hash table with corresponding counts for each token.
- The structure of this database reflects simple criteria:
  - ▣ A token column to represent each unique token
  - ▣ A column for a total count per token
  - ▣ A column for the total count per seed source (website)

# Token statistics

- The PsyCoL Maltese corpus...
  - ...has 3,323,325 total tokens.
  - ...has 53,396 unique tokens.
  - ...contains 58.9% web-crawled data.
  - ...contains 41.1% data from Dr. Albert Gatt.
  - ...represents the largest tokenized accessible corpus of the Maltese language.
- The PsyCoL Hebrew corpus...
  - ...has 60,052,261 tokens.
  - ...has 396,469 unique tokens.

# Corpus interface

- Goals of the interface:
  - Universal, free accessibility
  - Cross-platform accessibility
  - Requirement-free accessibility

<http://dingo.sbs.arizona.edu/~psycol/resources/>

# Tools

- The corpus interface currently provides several tools or calculators that provide useful lexical statistics for factors that have been shown to play important roles in lexical access:
  - Token frequency calculator
  - Neighborhood density calculator
  - Lexical uniqueness point calculator

# Why these tools?

- We care about frequency because in our psycholinguistic experiments, we manipulate frequency among our items.
- 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).
- Finally, the lexical uniqueness point of a given word plays a role in lexical access (e.g., Marslen-Wilson 1978, Wurm 2007).

# Why these tools?

- Having access to these three tools allows us to design more valuable experiments.
- Without access to such tools, it would be impossible to test the effects of these variables (frequency, density, uniqueness point) in our experiments.
- Of course, other tools would be useful as well, e.g., a morphological parser.

# Testing the corpora

- One way to assess corpus goodness is to compare how accurately it encodes these variables.
- Here, we chose to test frequency and compare it with subjective word familiarity in Maltese.

# Why word familiarity?

- Other researchers have established the role of word familiarity in lexical retrieval (Gernsbacher 1984, Connine et al. 1990, and many others).
- In general, the more familiar a word, the faster it is recognized.
- Familiarity is calculated via subjective familiarity judgments, which are provided by native speakers.

# How is familiarity different from frequency?

- In a ground-breaking study, Gernsbacher (1984) teased apart frequency and familiarity, and established that familiarity can play a more robust role than frequency in explaining differences and discrepancies she discovered in a number of earlier studies.
- In subsequent studies (Connine et al. 1990), the more robust nature of familiarity was borne out by the fact that its effects are obtained in both auditory and visual naming tasks, while frequency effects may not always be found in the auditory domain.

# Frequency and familiarity in Maltese

- No researcher has yet documented or performed quantitative analysis on word familiarity or word frequency for Maltese.
- This has mainly been due to the lack of a substantial corpus (for frequency) and the lack of subjective ratings (for familiarity).

# The data

- Because there is little previous quantitative research on Maltese, we chose a very narrow focus in our investigation: Semitic-origin verbs in Maltese.

# Background on Maltese verbs

- The Semitic-origin verbs of Maltese conform to the classical Semitic *binyan* system, in which verbs are organized into classes that share morphophonological, syntactic, and semantic properties.

# Maltese binyan system

<i>Binyan</i>	<i>Function</i>
1	<ul style="list-style-type: none"><li>• basic active (transitive or intransitive)</li></ul>
2	<ul style="list-style-type: none"><li>• intensive of 1</li><li>• transitive of 1</li></ul>
3	<ul style="list-style-type: none"><li>• transitive of 1</li></ul>
5	<ul style="list-style-type: none"><li>• passive of 2</li><li>• reflexive of 2</li></ul>
6	<ul style="list-style-type: none"><li>• passive of 2</li><li>• reflexive of 3</li></ul>
7	<ul style="list-style-type: none"><li>• passive of 1</li><li>• reflexive of 1</li></ul>
8	<ul style="list-style-type: none"><li>• passive of 1</li><li>• reflexive of 1</li></ul>
9	<ul style="list-style-type: none"><li>• inchoative, acquisition of a quality</li></ul>
10	<ul style="list-style-type: none"><li>• originally inchoative</li></ul>
Q1	<ul style="list-style-type: none"><li>• basic active</li></ul>
Q2	<ul style="list-style-type: none"><li>• passive and/or reflexive of Q1</li></ul>

# Research questions

- How categorical are the different verbal binyanim?
- Are some binyanim more familiar to native speakers than others?
- Are some binyanim more frequent than others?

# Experiment 2a: word familiarity

- In a word familiarity experiment, subjects are presented with a word and asked to rate how familiar the word is to them.
- Responses are then averaged across subjects to yield a mean word familiarity for each item and for each binyan.
- We then test whether some binyanim are significantly more familiar than others.

# Experiment 2a: word familiarity

- With the help of Dr. Albert Gatt, we designed and ran a subjective word familiarity experiment with all Semitic-origin Maltese verbs occurring in the Aquilina (1978) Maltese-English dictionary.
- The experiment was carried out online.
- 107 native Maltese speakers living in Malta participated in the experiment.

# Experiment 2a: procedure

- In the experiment, each subject logged into a secure website.
- The experiment began by instructing the subject to rate each verb with respect to how familiar they were with the verb.
- The rating was indicated via a slider bar, whose leftmost edge corresponded to “not familiar to me at all” and whose rightmost edge corresponded to “very familiar to me”.

# Experiment 2a: procedure

Note: these instructions are translated into English here, but were provided in Maltese for the actual experiment.

From now on, you're going to see a number of Maltese verbs. Below each one, you'll see a slider like the one you can see below.

[slider appears on next slide so you can see it better]

For each verb that you see, move the slider to the right to reflect **how familiar that verb is to you**. Your judgement should be based on your own, subjective impression as a speaker of Maltese.

**The further to the right you move the slider, the more the verb is familiar to you.** Leaving the slider at its initial position, without moving it, means that this verb is completely unfamiliar to you.

There are no correct or incorrect responses; the most important thing is what you feel as a speaker of Maltese.

It is important that you remember that all the words you will be reading are *verbs*. In order to help you remember this, every word will appear in the context of the pronoun *huwa* (e.g. *huwa ftakar*, *huwa ttiekkel*).


After you've judged the familiarity of a word, you can continue by clicking on the *kompli* ("continue") button. If you wish to stop, just click the *ieqaf hawn* ("stop here") button.

# Experiment 2a: word familiarity slider

The rating slider:

Agħti ġudizzju ta' kemm huwa familjari għalik dan il-verb:

(Huwa) hareġ



kompli   ieqaf hawn

# Experiment 2a: procedure

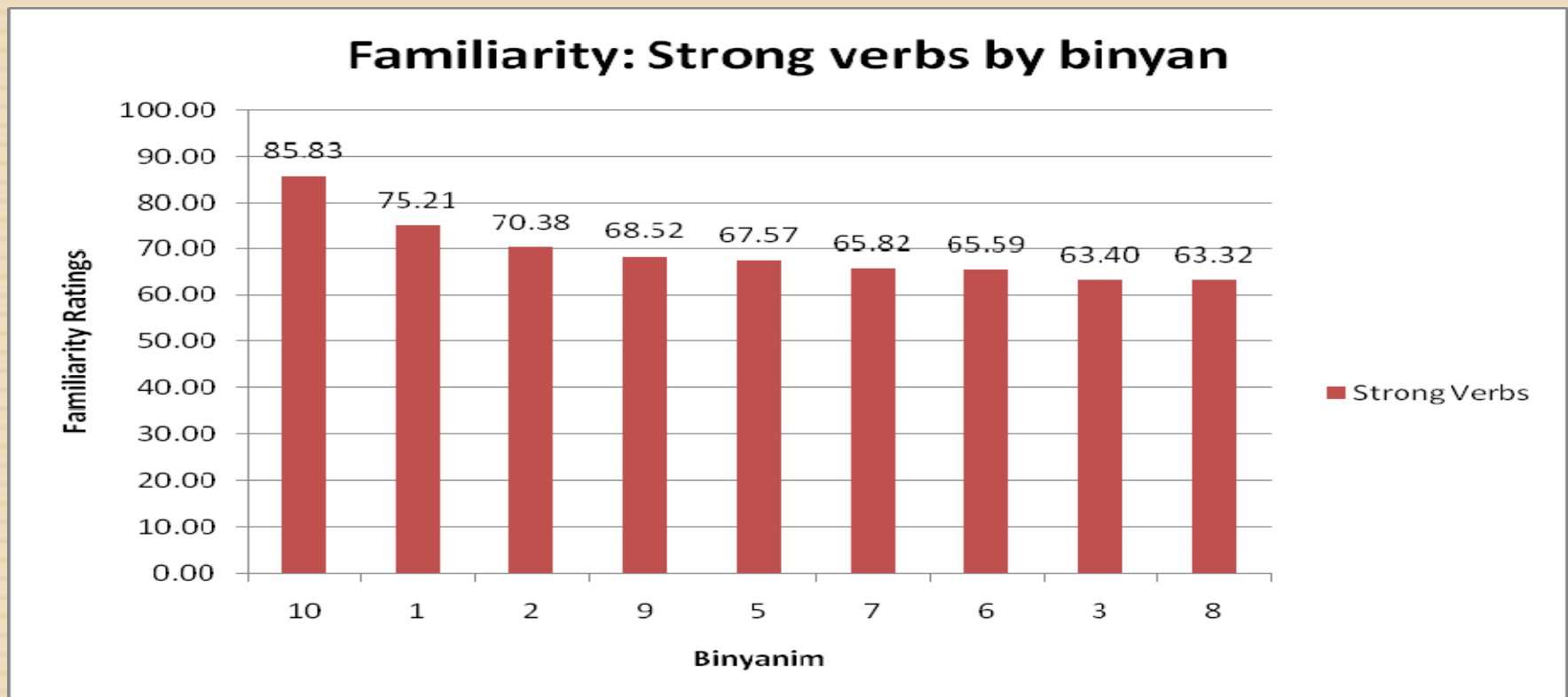
- Because our experiment included a large number of verbs (1536 verbs total), we did not require each subject to rate each and every verb.
- Instead, each subject was given the choice of exiting the experiment at any point.
- The unrated items from one subject were then placed at the top of the list of randomized items presented to the next subject.

# Experiment 2a: results

- In all, 107 subjects rated a total of 1536 verbs.
- Each verb was rated an average of 6.01 times.
- Each subject rated an average of 55.15 verbs.

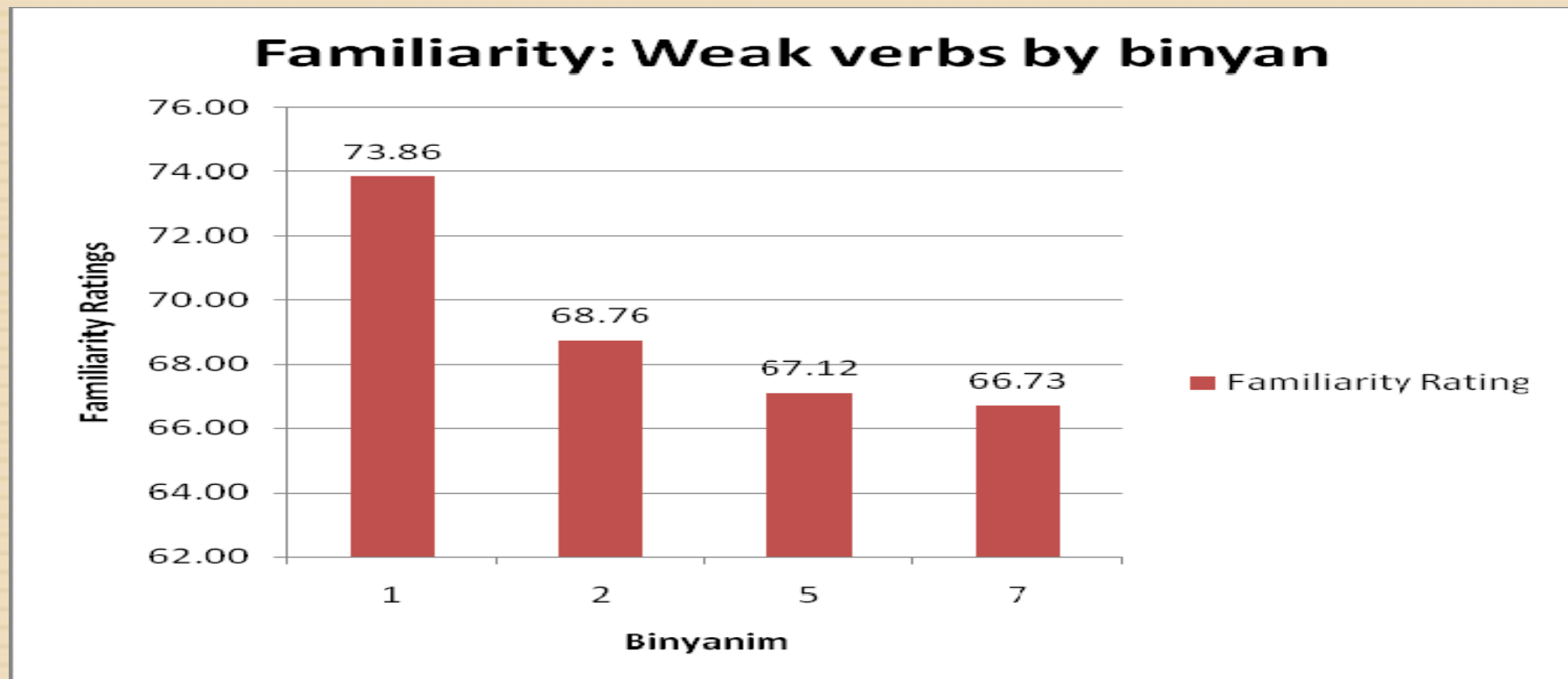
# Experiment 2a: results

- For strong verbs



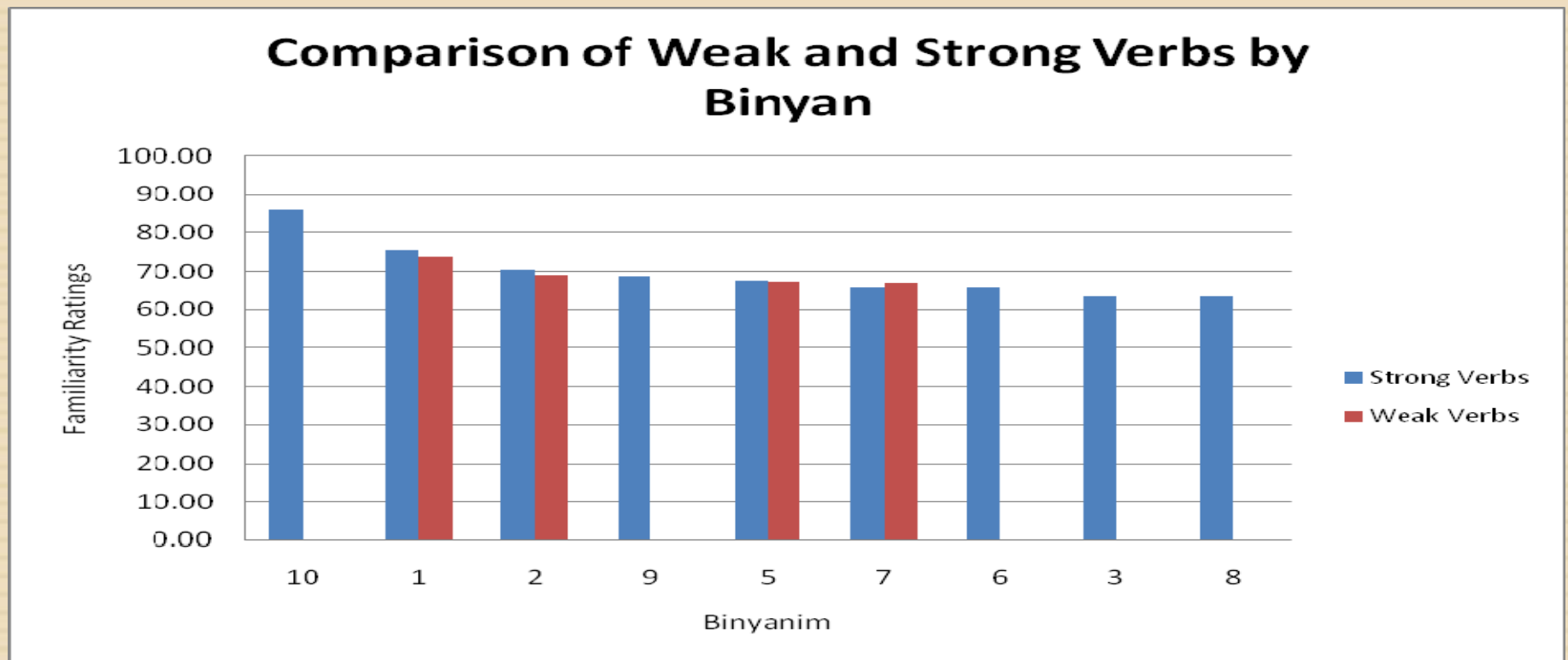
# Experiment 2a: results

- For weak verbs



# Experiment 2a: results

- Strong and weak verbs together



# Which word familiarity differences are significant?

## □ Strong verbs:

□ Binyan 1 is more familiar than Binyan 5:  $t(436)=2.935, p=0.004$

□ Binyan 1 is more familiar than Binyan 8:  $t(202)=1.963, p=0.051$

## □ Strong+weak combined:

□ Binyan 1 is more familiar than Binyan 2:  $t(597)=2.124, p=0.034$

□ Binyan 1 is more familiar than Binyan 5:  $t(533)=3.06, p=0.002$

□ Binyan 1 is more familiar than Binyan 7:  $t(280.526)=2.904, p=0.004$

# Experiment 2b: word frequency

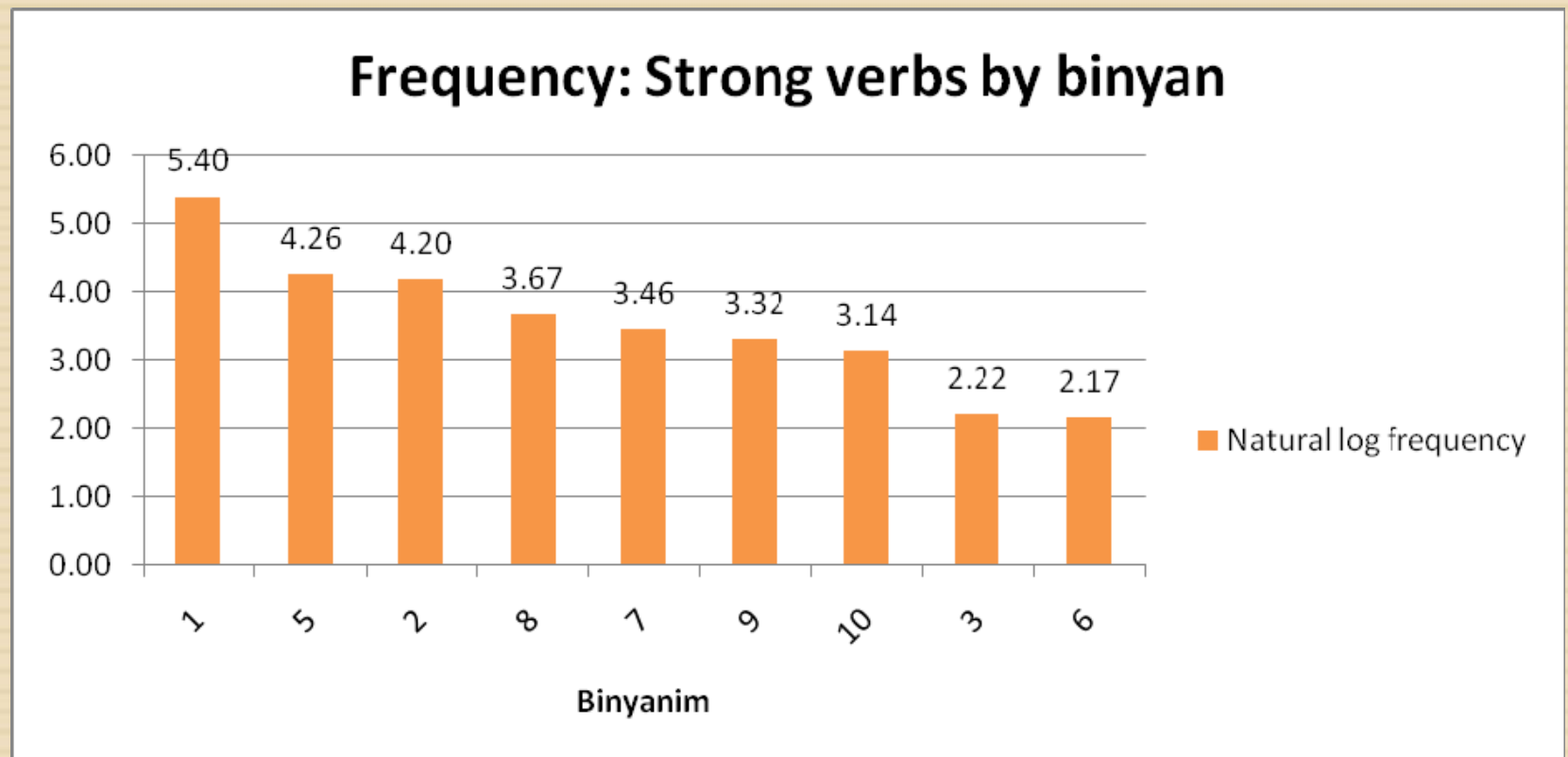
- As a comparison, we also measured Maltese word frequency.
- Word frequency was calculated using our corpus.

# Experiment 2b: methodology

- Word frequency can be calculated using regular expressions in the PsyCoL Maltese Lexical Corpus.
- In order to include inflected forms of verb stems, our regular expressions contained prefixed and suffixed versions of all Semitic-origin verb stems taken from Aquilina (1978).
- Frequency is calculated in terms of raw frequency (the corpus has over 3 million items), frequency per million, and log frequency.
- Results here are reported in log notation, since in lexical access, the effect of frequency is logarithmic (and not linear).

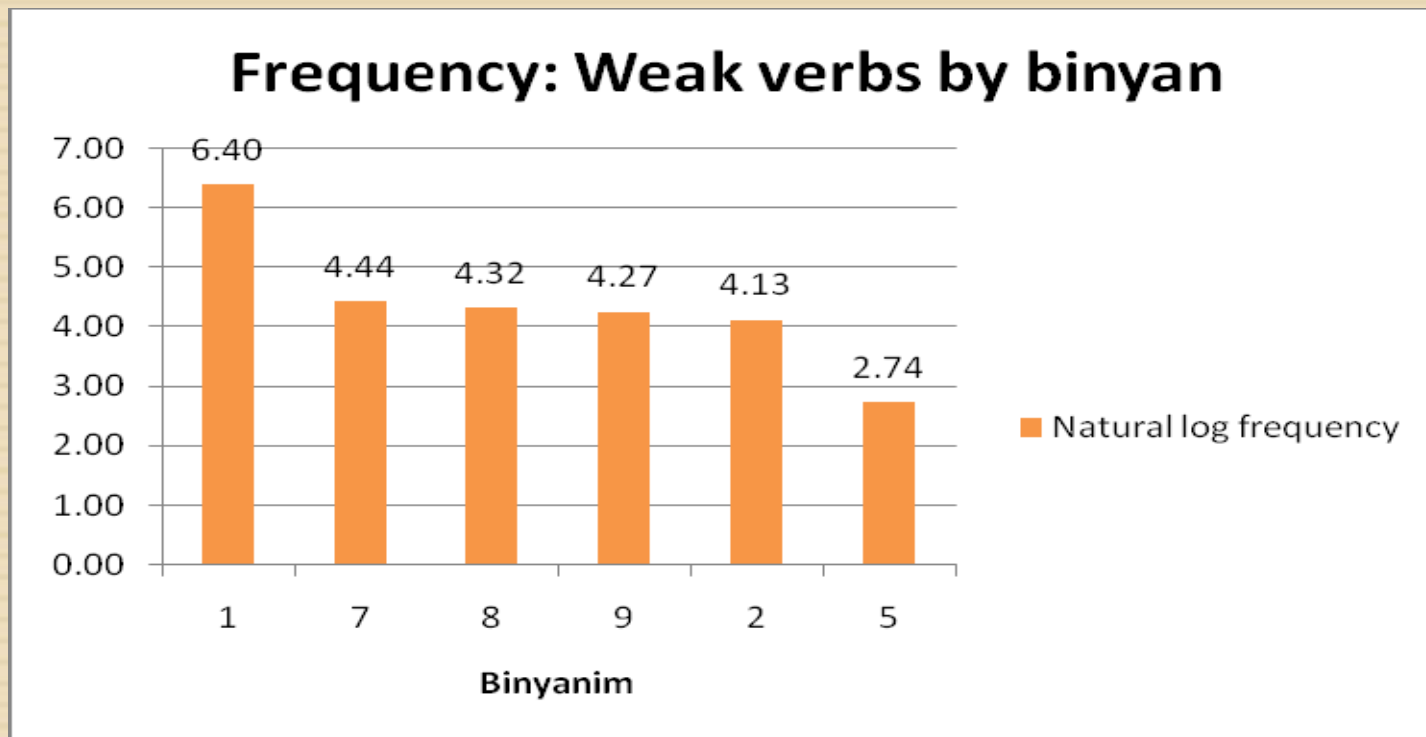
# Experiment 2b: results

□ For strong verbs:



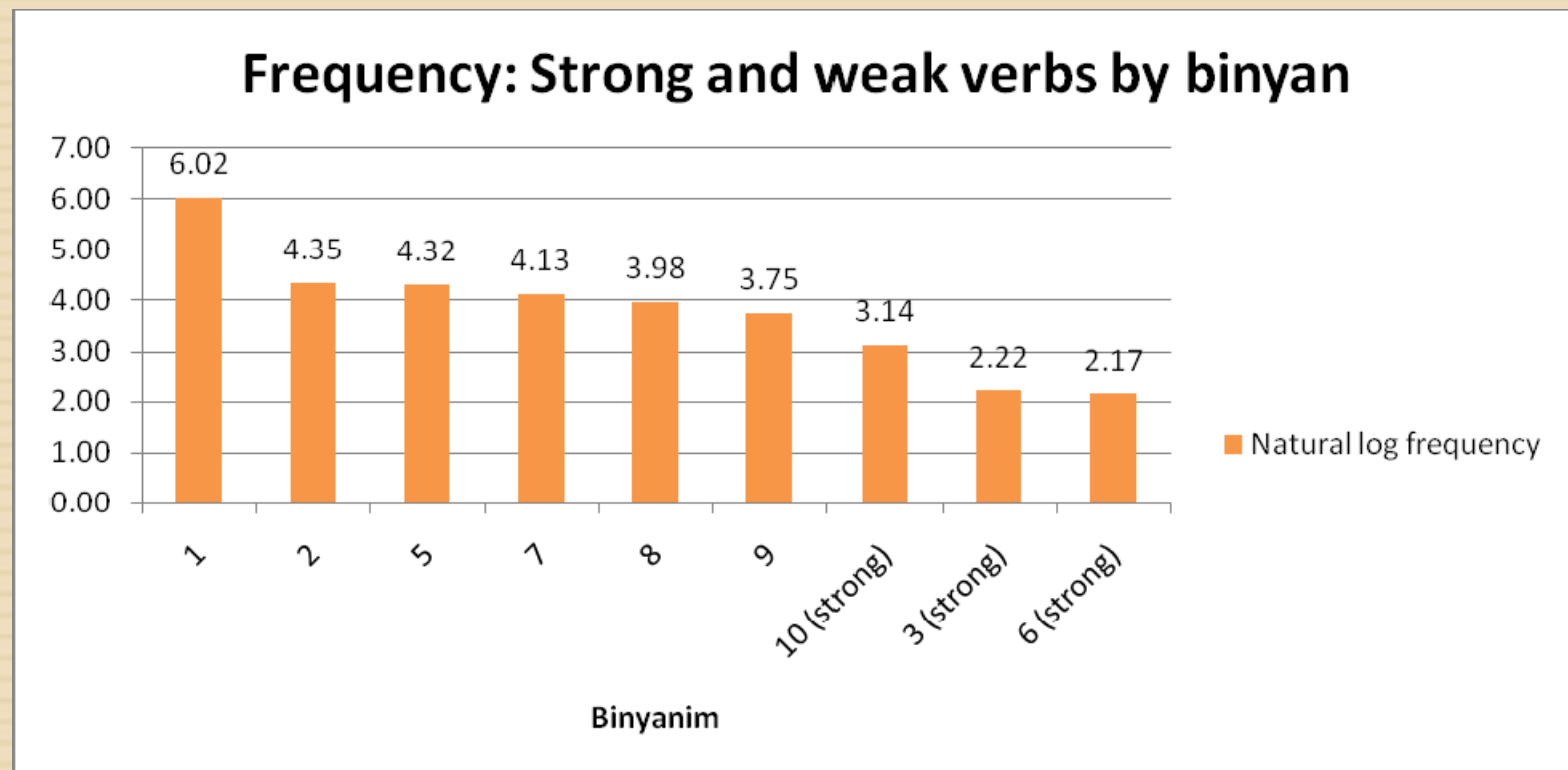
# Experiment 2b: results

□ For weak verbs:



# Experiment 2b: results

- Strong and weak verbs together:



# Which word frequency results are significant?

## □ Strong verbs:

□ Binyan 1 is more frequent than Binyan 5:  $t(150.288)=2.448, p = 0.016$

## □ Weak verbs:

□ Binyan 1 is more frequent than Binyan 2:  $t(81)=2.96, p=0.004$

□ Binyan 1 is more frequent than Binyan 5:  $t(69)=3.545, p=0.001$

□ Binyan 8 is more frequent than Binyan 5:  $t(29)=2.821, p=0.009$

# Which word frequency results are significant?

## □ Strong and weak together:

- Binyan 1 is more frequent than Binyan 2:  $t(228.195) = 3.494, p=0.001$
- Binyan 1 is more frequent than Binyan 3:  $t(9.416) = 3.048, p=0.013$
- Binyan 1 is more frequent than Binyan 5:  $t(222.992) = 4.37, p=0.000$
- Binyan 5 is more frequent than Binyan 9:  $t(122) = 2.266, p=0.025$

# Corpus linguistics and psycholinguistics

- Maltese binyanim do cluster together as far as frequency and familiarity are concerned, but not all the differences are significant.
- In other words, some binyanim are significantly more frequent than others, and some binyanim are significantly more familiar than others.
  - ▣ Frequency:  $1 > 2$ ,  $1 > 3$ ,  $1 > 5$ ,  $5 > 9$  (for both strong and weak)
  - ▣ Familiarity:  $1 > 2$ ,  $1 > 5$ ,  $1 > 7$  (for both strong and weak)
- Across the two experiments, two of the same comparisons turned out to be significant.
- These specific data can be used for a number of purposes, including for designing psycholinguistic experiments in Maltese that manipulate frequency and/or familiarity.

# Future work

- How do Hebrew binyanim look?
  - We have recently collected word familiarity data on Hebrew, but word frequency data are difficult to obtain given the variable nature of the orthography in our corpus.
- What other lexical access studies might shed light on root-based vs. word-based storage?
  - Priming studies (including subliminal speech priming)
  - Studies using pseudo-words formed from existing roots

# Conclusion

- The intersection of corpus linguistics and psycholinguistics can contribute to our understanding of the nature of words and the organization of the lexicon.
- Under-documented languages represent a useful resource since they may provide an unusual confluence of properties (orthographic and morphological).
- Corpus creation for such a language is a worthwhile and straightforward investment, and can provide the chance to design experiments that avoid confounds found in better-studied languages.

# Thank you!

- Many thanks for your attention.
- As always, your questions, comments, and feedback are welcome 😊
  - [ussishki@u.arizona.edu](mailto:ussishki@u.arizona.edu)