Creating a conversational Hebrew vocabulary list

A reproducible use of technology to overcome scarcity of data

Juan D. Pinto

April 22, 2018

Overview

- Why a frequency dictionary?
- What is it?
- How is it created?
- Challenges

WHY

Research applications

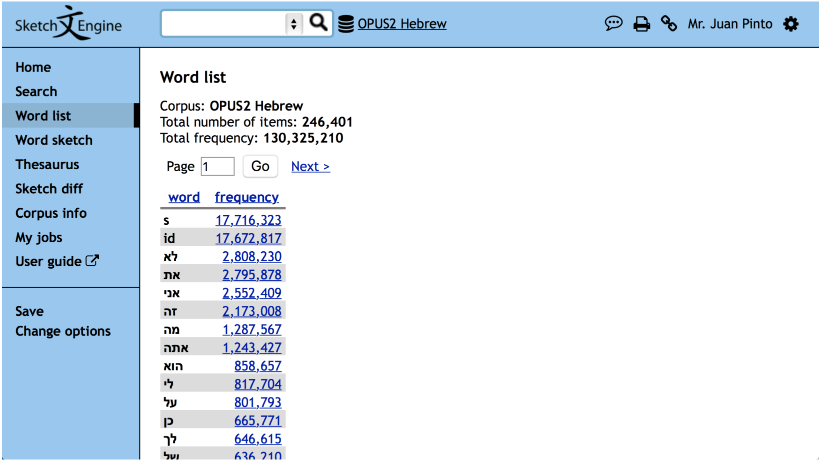
- Traditional linguistic studies that look for common morphological patterns
- Corpus-linguistic studies seeking to understand language through “real world” texts
- Psycholinguistic studies that explore connections between a speaker’s mental lexicon and word frequency

Practical applications

- Curriculum and textbook planning (prioritizing vocabulary)
- Vocabulary selection for graded readers and dictionaries
- Independent language study
- Calculating a text's vocabulary load

- How can vocabulary knowledge be appropriately tested and measured?
- What is the role of extensive reading (as opposed to intensive reading) in incidental vocabulary acquisition?
- What level of vocabulary do learners need in order to read extensively for pleasure?
- What level of vocabulary do learners need in order to succeed in an academic setting?
- What role does specialized vocabulary play in reaching understanding?

Sketch Engine: *https://www.sketchengine.eu*



Word list

Corpus: **OPUS2 Hebrew**
Total number of items: **246,401**
Total frequency: **130,325,210**

Page 1 Go Next >

word	frequency
s	17,716,323
id	17,672,817
לא	2,808,230
את	2,795,878
אני	2,552,409
זה	2,173,008
מה	1,287,567
אתה	1,243,427
הוא	858,657
לי	817,704
על	801,793
כן	665,771
לך	646,615
של	636,210

What

The Conversational Hebrew Vocabulary List (CHVL)

OPUS-frequencies repository:
https://github.com/juandpinto/opus-frequencies

What is a ***word***?

1. **Token =** total number of words

What is a *word*?

1. **Token =** total number of words
2. **Type =** number of separate and distinct words

What is a *word*?

1. **Token =** total number of words
2. **Type =** number of separate and distinct words
3. **Lemma =** "A set of lexical forms having the same stem and belonging to the same major word class, differing only in inflection and/or spelling." (Francis, Kučera, & Mackie, 1982)

What is a *word*?

1. **Token =** total number of words
2. **Type =** number of separate and distinct words
3. **Lemma =** "A set of lexical forms having the same stem and belonging to the same major word class, differing only in inflection and/or spelling." (Francis, Kučera, & Mackie, 1982)
4. **Word family =** English taxonomy by Bauer and Nation (1993).

Frequency and range

- **Raw frequency =** total number of times that the word is attested in the corpus

Frequency and range

- **Raw frequency =** total number of times that the word is attested in the corpus
- **Normalized frequency =** how many times the word appears *for every x tokens* in the corpus

Frequency and range

- **Raw frequency =** total number of times that the word is attested in the corpus
- **Normalized frequency =** how many times the word appears *for every x tokens* in the corpus
- **Range =** the number of sub-corpora in which the word can be found

Dispersion

- Dispersion (simplistic) = combination of both frequency and range

Dispersion

- Dispersion (simplistic) = combination of both frequency and range
    - Incorporates benefits of both

Dispersion

- Dispersion (simplistic) = combination of both frequency and range
    - Incorporates benefits of both
- U_{DP} = usage coefficient of Gries' *deviation of proportions*, or *DP* (Gries, 2008; 2010)

Dispersion

- Dispersion (simplistic) = combination of both frequency and range
  - Incorporates benefits of both
- U_{DP} = usage coefficient of Gries' *deviation of proportions*, or *DP* (Gries, 2008; 2010)
  - *DP* x *frequency* (Matsushita, 2012, p. 99; Sorell, 2013, p. 89)

$$\mathrm{U}_{\mathrm{DP}} =$$

$$\left(1 - 0.5\sum_{i=1}^{n}\left|\frac{file_i\ tokens}{total\ tokens} - \frac{frequency_x\ in\ file_i}{total\ frequency_x}\right|\right) \times total\ frequency_x$$

HOW

A.K.A. *Methods*

Python

1. Along with R, Python is one of the most widely-used languages for this type of data analysis

Python

1. Along with R, Python is one of the most widely-used languages for this type of data analysis
2. Python was designed specifically to be very readable and easy to learn

Python

1. Along with R, Python is one of the most widely-used languages for this type of data analysis
2. Python was designed specifically to be very readable and easy to learn
   - Easy to understand the syntax

Python

1. Along with R, Python is one of the most widely-used languages for this type of data analysis
2. Python was designed specifically to be very readable and easy to learn
    - Easy to understand the syntax
    - Widely considered good for beginners because of its simplicity

Steps

1. Find a corpus
2. Clean the corpus
3. Extract data
4. Make calculations
5. Sort and export

1. Find a corpus

OpenSubtitles2018

OPUS: *http://opus.nlpl.eu*

Parsed corpus example

```
<s id="49">
  <time value="00:03:22,280" id="T39S" />
  <w xpos="ADV" head="49.3" feats="PronType=Int" upos="ADV"
      id="49.1" deprel="obj"> </w>
  <w xpos="PRON" head="49.3" feats="Gender=Masc|Number=Sing
      PronType=Prs" upos="PRON" lemma="  " id="49.2" deprel
  <w xpos="VERB" head="0" feats="Gender=Masc|HebBinyan=PAAL
      Person=1,2,3|VerbForm=Part|Voice=Act" upos="VERB" mis
      lemma="  " id="49.3" deprel="root">  </w>
  <w xpos="PUNCT" head="49.3" upos="PUNCT" lemma="," id="49
      deprel="punct">,</w>
  <w xpos="NOUN" head="49.3" feats="Gender=Masc|Number=Sing
      misc="SpaceAfter=No" lemma="   " id="49.5" deprel="ob
  <w xpos="PUNCT" head="49.3" upos="PUNCT" misc="SpaceAfte
      id="49.6" deprel="punct">?</w>
  <time value="00:03:24,120" id="T39E" />
</s>
```

2. Clean the corpus

```
Zipped folder in GZ format
    Folder for year X
        Folder for movie A
            Zipped XML in GZ format
            Zipped XML in GZ format
            Zipped XML in GZ format
        Folder for movie B
            Zipped XML in GZ format
            Zipped XML in GZ format
    Folder for year Y
        Folder for movie C
            Zipped XML in GZ format
        Folder for movie D
            Zipped XML in GZ format
            Zipped XML in GZ format
            Zipped XML in GZ format
        Folder for movie E
            Zipped XML in GZ format
            Zipped XML in GZ format
```

3. Extract data

```
' ': {
          '/he/0/5753574/6853341.xml': 168,
          '/he/0/3607000/5764778.xml': 94},
'  ': {
          '/he/0/5753574/6853341.xml': 3},
'   ': {
          '/he/0/5753574/6853341.xml': 6,
          '/he/0/3607000/5764778.xml': 2,
          '/he/0/1278351/3777598.xml': 1}
```

4. Make calculations

Normalized frequency

$$\left(\frac{raw\ frequency}{total\ frequency}\right) \times\ 1,000,000$$

4. Make calculations

Range

4. Make calculations

Dispersion (U_{DP})

$$\left(1-0.5\sum_{i=1}^{n}\left|\frac{file_i\ tokens}{total\ tokens}-\frac{frequency_x\ in\ file_i}{total\ frequency_x}\right|\right)\times total\ frequency_x$$

5. Sort and export

Challenges

- Ideal vs. existing corpus

Challenges

- Ideal vs. existing corpus
    - Very accessible and cost effective

Challenges

- Ideal vs. existing corpus
  - Very accessible and cost effective
  - High correlation between subtitles and conversational language (Brysbaert & New, 2009; New et al., 2007)

Challenges

- Ideal vs. existing corpus
  - Very accessible and cost effective
  - High correlation between subtitles and conversational language (Brysbaert & New, 2009; New et al., 2007)
- Translated vs. original language

Challenges

- Ideal vs. existing corpus
  - Very accessible and cost effective
  - High correlation between subtitles and conversational language (Brysbaert & New, 2009; New et al., 2007)
- Translated vs. original language
  - Quantity varies

Challenges

- Ideal vs. existing corpus
  - Very accessible and cost effective
  - High correlation between subtitles and conversational language (Brysbaert & New, 2009; New et al., 2007)
- Translated vs. original language
  - Quantity varies
- Automatic parser

Conclusions

- Frequency dictionaries are useful for education and research

Conclusions

- Frequency dictionaries are useful for education and research
- Many tools

Conclusions

- Frequency dictionaries are useful for education and research
- Many tools
- Minimal, simple coding helps

Don't leave all the fun to the English researchers!