

Hebrew Statistical Linguistics using a Morphologically Analyzed Blog Corpus

Tal Linzen

Linguistics Department
Tel Aviv University

Israeli Seminar on Computational Linguistics

June 16, 2010

- 1 Motivation
- 2 The corpus
- 3 Use case 1: Possessive Datives
- 4 Use case 2: Verb frame bias
- 5 Conclusion and sermon

Corpus linguistics

A very short introduction

- In the American structuralist school, corpora were the only way to do linguistics (e.g. Bloomfield's *Language*)

Corpus linguistics

A very short introduction

- In the American structuralist school, corpora were the only way to do linguistics (e.g. Bloomfield's *Language*)
- The Chomskyan revolution questioned the value of corpora

Corpus linguistics

A very short introduction

- In the American structuralist school, corpora were the only way to do linguistics (e.g. Bloomfield's *Language*)
- The Chomskyan revolution questioned the value of corpora
- Corpora are experiencing a renaissance, partly due to the abundance of large electronic corpora

Corpus linguistics

A very short introduction

- In the American structuralist school, corpora were the only way to do linguistics (e.g. Bloomfield's *Language*)
- The Chomskyan revolution questioned the value of corpora
- Corpora are experiencing a renaissance, partly due to the abundance of large electronic corpora
- Two kinds of corpus linguistics:

Corpus linguistics

A very short introduction

- In the American structuralist school, corpora were the only way to do linguistics (e.g. Bloomfield's *Language*)
- The Chomskyan revolution questioned the value of corpora
- Corpora are experiencing a renaissance, partly due to the abundance of large electronic corpora
- Two kinds of corpus linguistics:
 - ① Getting real examples for qualitative analyses

Corpus linguistics

A very short introduction

- In the American structuralist school, corpora were the only way to do linguistics (e.g. Bloomfield's *Language*)
- The Chomskyan revolution questioned the value of corpora
- Corpora are experiencing a renaissance, partly due to the abundance of large electronic corpora
- Two kinds of corpus linguistics:
 - ① Getting real examples for qualitative analyses
 - ② Quantitative analyses (“**Statistical linguistics**”)

Why not use Google?

- Google are doing a pretty good job indexing the web, do we need another corpus?

Why not use Google?

- Google are doing a pretty good job indexing the web, do we need another corpus?
- **Yes:** Google provides very limited search facilities, especially for languages with complex morphology (such as Hebrew)

Why not use Google?

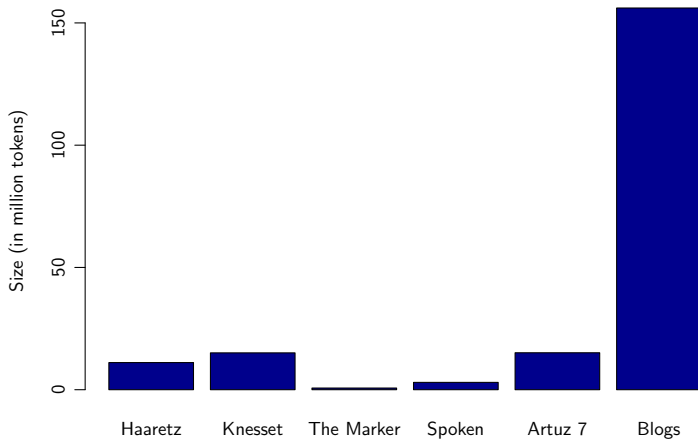
- Google are doing a pretty good job indexing the web, do we need another corpus?
- **Yes:** Google provides very limited search facilities, especially for languages with complex morphology (such as Hebrew)
- Results are not reproducible and counts are unreliable – unsuitable for Type 2 corpus linguistics

Why not use Google?

- Google are doing a pretty good job indexing the web, do we need another corpus?
- **Yes:** Google provides very limited search facilities, especially for languages with complex morphology (such as Hebrew)
- Results are not reproducible and counts are unreliable – unsuitable for Type 2 corpus linguistics
- Still, good source of ideas and examples for Type 1 corpus linguistics of (which is always a part of Type 2 work)

Why another corpus?

Available Hebrew corpora



Israblog

The screenshot shows the nana10.co.il website. The browser address bar displays the URL `http://israblog.nana10.co.il/`. The page header includes the nana10 logo and the text "החיים זה כאן". Below the header is a search bar and a navigation menu with links like "Getting Started", "Latest Headlines", "The LINGUIST List", and "Facebook | Home".

The main content area features a large blue banner with the text "אמא, מידע זה חשוב עבור בתך" (Mom, this information is important for your daughter) and a photo of two women. Below this banner is a navigation bar with links: "בלוגים לפי קטגוריות", "סוגות בלוגים", "בלוגים פעילים במיוחד", "בלוג אקטיוו", "חזרה לרף הבית הקלאסי".

The page is divided into several columns:

- Left Sidebar:** A vertical advertisement for "Maka" with the text "הית פעם עם ברשית אמריקאית או נערת הוואי".
- Main Content Grid:** A grid of news items with images and titles:
 - סוגים חדשים:** כולל תמונה של ילד.
 - אמרי:** כולל תמונה של אדם יושב על הרצפה.
 - שיעורי בית:** כולל תמונה של אדם לומד.
 - המלחה סקנדלית:** כולל תמונה של כדור הארץ.
- Right Sidebar:**
 - פוסט למערכת:** Form for submitting posts.
 - למתיירות בלוג חדש:** Section for new bloggers.
 - אמא, מידע זה חשוב עבור בתך:** A smaller version of the main banner.
 - מסדרים:** Section for orders.

The footer of the browser window shows a search bar with "Find" and "Done" buttons, and a "Proxy Name" field.

Technicalities

- Limiting data acquisition to a single site simplifies HTML cleanup (“scraping”)

Technicalities

- Limiting data acquisition to a single site simplifies HTML cleanup (“scraping”)
- Site has a handy “random user” feature, no need to work hard to discover users

Technicalities

- Limiting data acquisition to a single site simplifies HTML cleanup (“scraping”)
- Site has a handy “random user” feature, no need to work hard to discover users
- Morphologically analyzed using Adler and Elhadad’s BGUTagger

Technicalities

- Limiting data acquisition to a single site simplifies HTML cleanup (“scraping”)
- Site has a handy “random user” feature, no need to work hard to discover users
- Morphologically analyzed using Adler and Elhadad’s BGUTagger
- No parser, so not parsed (hint!)

Properties of the corpus

- Clearly marked for author – useful for exploring (or controlling for) individual variation

Properties of the corpus

- Clearly marked for author – useful for exploring (or controlling for) individual variation
- Users can specify their age and gender, and most do

Properties of the corpus

- Clearly marked for author – useful for exploring (or controlling for) individual variation
- Users can specify their age and gender, and most do
- Variety of ages and styles, not copy-edited

Properties of the corpus

- Clearly marked for author – useful for exploring (or controlling for) individual variation
- Users can specify their age and gender, and most do
- Variety of ages and styles, not copy-edited
- On the other hand, not balanced

Use case 1: Possessive Datives

- **Possessive Dative:**

- (1) šavarti le-šaul et ha-kos.
I.broke to-Shaul ACC the-glass
'I broke Shaul's glass.' (*contested gloss*)

Use case 1: Possessive Datives

- **Possessive Dative:**

- (1) šavarti le-šaul et ha-kos.
I.broke to-Shaul ACC the-glass
'I broke Shaul's glass.' (*contested gloss*)

- **Ordinary possession:**

- (2) šavarti et ha-kos šel šaul.
I.broke ACC the-glass of Shaul
'I broke Shaul's glass.'

Possessive Datives: Hypotheses

- Based on the behavior of similar constructions in other languages, we expect that:

Possessive Datives: Hypotheses

- Based on the behavior of similar constructions in other languages, we expect that:
 - **Synchronic:** the Possessive Dative should be more popular when the possessed object is a body part

Possessive Datives: Hypotheses

- Based on the behavior of similar constructions in other languages, we expect that:
 - **Synchronic:** the Possessive Dative should be more popular when the possessed object is a body part
 - **Diachronic:** this preference should diminish with time

Possessive Datives: Annotation

- The dative preposition *le* 'to' is fused with the word in Hebrew orthography, so in general it's hard to search for dative sentences

Possessive Datives: Annotation

- The dative preposition *le* 'to' is fused with the word in Hebrew orthography, so in general it's hard to search for dative sentences
- Luckily, corpus is morphologically analyzed

Possessive Datives: Annotation

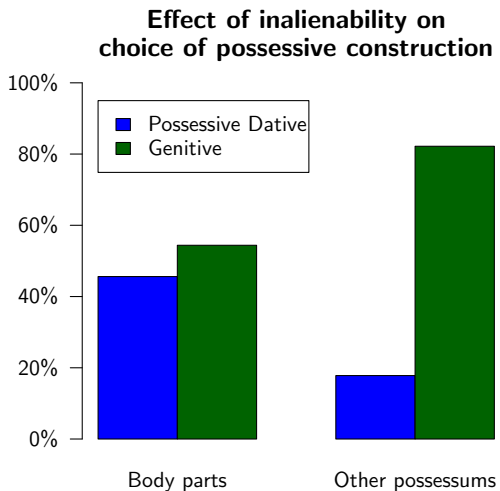
- The dative preposition *le* 'to' is fused with the word in Hebrew orthography, so in general it's hard to search for dative sentences
- Luckily, corpus is morphologically analyzed
- Dative constructions with common transfer verbs (*give* etc.) were automatically removed

Possessive Datives: Annotation

- The dative preposition *le* 'to' is fused with the word in Hebrew orthography, so in general it's hard to search for dative sentences
- Luckily, corpus is morphologically analyzed
- Dative constructions with common transfer verbs (*give* etc.) were automatically removed
- The remaining sentences were filtered manually, to remove irrelevant uses of the preposition *le*

Possessive Datives: The synchronic hypothesis

Linzen (2009)



Possessive Datives: The diachronic hypothesis

- Reminder: the preference for body parts should diminish with time

Possessive Datives: The diachronic hypothesis

- Reminder: the preference for body parts should diminish with time
- **Problem:** Corpus is not historical

Possessive Datives: The diachronic hypothesis

- Reminder: the preference for body parts should diminish with time
- **Problem:** Corpus is not historical
- Solution: use **ages** reported by bloggers

Possessive Datives: The diachronic hypothesis

- Reminder: the preference for body parts should diminish with time
- **Problem:** Corpus is not historical
- Solution: use **ages** reported by bloggers
- Analyzed using a mixed-effects logistic regression model

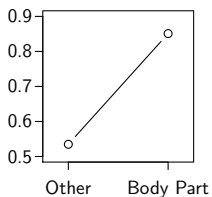
Regression coefficients

Factor	Estimate	Std. Error	z value	Pr	Sig
(Intercept)	0.507	0.101	5.035	0.000	***
17-18	-0.089	0.055	-1.609	0.108	
19-21	-0.094	0.057	-1.661	0.097	.
22-26	-0.051	0.091	-0.560	0.575	
27-35	-0.144	0.104	-1.394	0.163	
36-60	-0.368	0.124	-2.976	0.003	**
bodypart	0.605	0.116	5.204	0.000	***
male	-0.294	0.152	-1.928	0.054	.
17-18:bodypart	0.135	0.098	1.375	0.169	
19-21:bodypart	0.195	0.100	1.953	0.051	.
22-26:bodypart	0.625	0.172	3.642	0.000	***
27-35:bodypart	0.741	0.192	3.856	0.000	***
36-60:bodypart	0.997	0.237	4.205	0.000	***

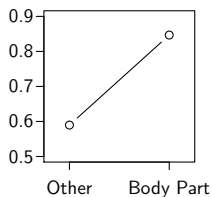
Age \times inalienability interaction

Effect of possessum inalienability on the probability of choosing the Possessive Dative increases with age

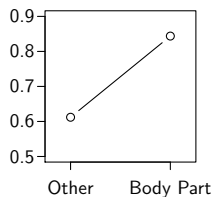
36-60



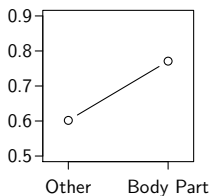
27-35



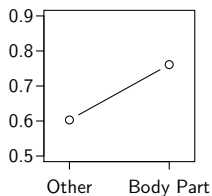
22-26



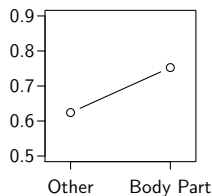
19-21



17-18



13-16



Use case 2: Verb frame bias

With Einat Shetreet and Naama Freedman

- fMRI experiment on the effect of frequency on language processing in the brain

Use case 2: Verb frame bias

With Einat Shetreet and Naama Freedman

- fMRI experiment on the effect of frequency on language processing in the brain
- Verbs can appear in more than one *frame* (syntactic context):
 - (3) I want an icecream.
 - (4) I want to eat.

Use case 2: Verb frame bias

With Einat Shetreet and Naama Freedman

- fMRI experiment on the effect of frequency on language processing in the brain
- Verbs can appear in more than one *frame* (syntactic context):
 - (3) I want an icecream.
 - (4) I want to eat.
- Compare three types of verbs:
 - ① Single frame
 - ② Multiple frames, frequency bias to one frame
 - ③ Multiple frames, no one frame takes a large portion of the frequency cake

Other applications of the corpus

- Word frequency norms for psycholinguistic experiments (in Naama Friedmann's lab)

Other applications of the corpus

- Word frequency norms for psycholinguistic experiments (in Naama Friedmann's lab)
- Hillel Taub-Tabib's corpus study on subject-verb inversion

Other applications of the corpus

- Word frequency norms for psycholinguistic experiments (in Naama Friedmann's lab)
- Hillel Taub-Tabib's corpus study on subject-verb inversion
- Nurit Melnik's research

Other applications of the corpus

- Word frequency norms for psycholinguistic experiments (in Naama Friedmann's lab)
- Hillel Taub-Tabib's corpus study on subject-verb inversion
- Nurit Melnik's research
- NLP development in Ben Gurion University

Conclusion and sermon

- Statistical linguistics requires large analyzed corpora

Conclusion and sermon

- Statistical linguistics requires large analyzed corpora
- The state of Hebrew NLP leaves much to be desired: morphological disambiguator can be improved, no parser

Thank you!