16. Linguistic corpora LING 471

Updates

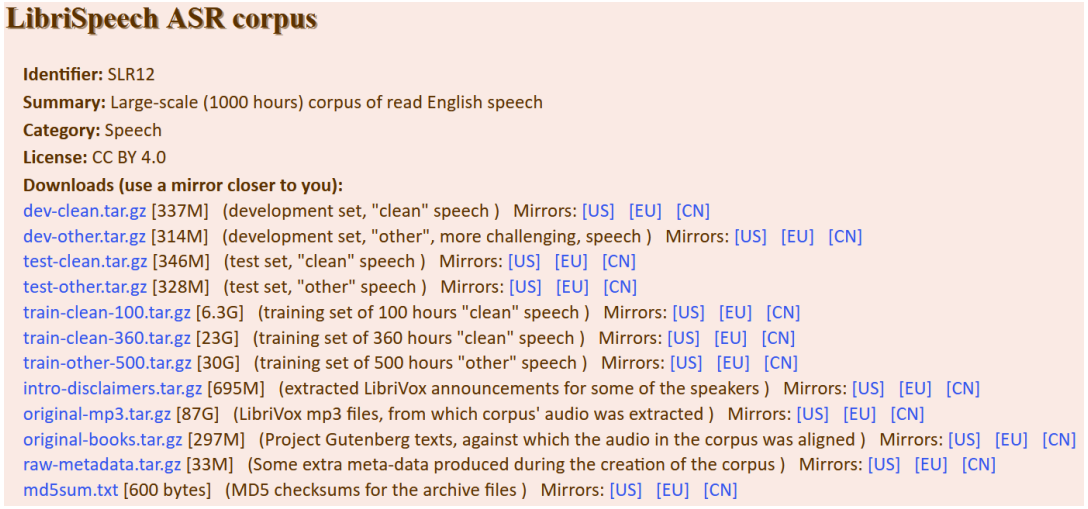
- Assignment 5 posted, due June 4
- Sample finals will be posted by the end of the week
- Remember to reply on Canvas discussion board for final presentation by this week
  - And check the readings before the presentation day as well

Learning outcomes

- Identify and define what a **corpus** is
- Describe what **annotations** look like for **text and speech corpora**
- Perform some **basic searches** in web interfaces for corpora
- Describe how **corpora** can be used to **study language**

What is a corpus?

- A **corpus** in linguistics is a large collection of data
  - Formal plural: **corpora**
- For textual data, a large collection of text documents
  - "Documents" in a general sense
  - Can include books, tweets, websites, YouTube comments, newspaper articles, etc.
- For speech data, a large collection of recordings
  - Usually needs to have some kind of annotation provided with it



Some common corpora

What do you do with corpora?

TEXTUAL

- Linguistic analysis
  - Find certain phrases; need to use RegEx
  - Can use some graphical tools for this (e.g., AntConc), sometimes requires programming
- Machine learning
  - E.g., assignments 2-5
  - Requires programming

SPEECH

- Linguistic analysis
  - Perform acoustic measurements on recordings
  - Can use some graphical tools (e.g., Praat), often requires programming
- Machine learning
  - E.g., speech recognition
  - Requires programming

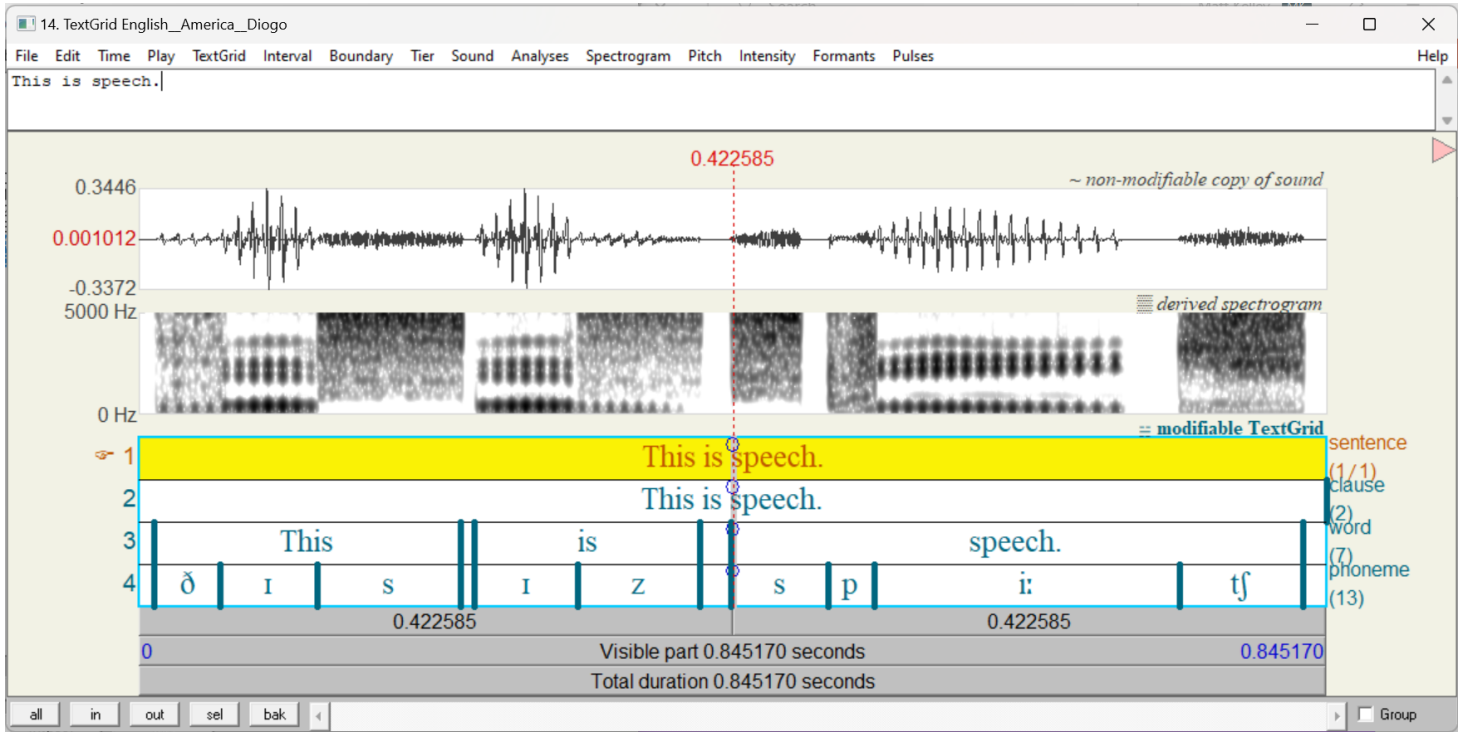
Battle-tested/NNP*/JJ industrial/JJ managers/NNS here/RB
always/RB buck/VB*/VBP up/IN*/RP nervous/JJ newcomers/NNS with/IN
the/DT tale/NN of/IN the/DT first/JJ of/IN their/PP$ countrymen/NNS to/TO
visit/VB Mexico/NNP ,/, a/DT boatload/NN of/IN samurai/NNS*/FW
warriors/NNS blown/VBN ashore/RB 375/CD years/NNS ago/RB ./.
"/" From/IN the/DT beginning/NN ,/, it/PRP took/VBD a/DT man/NN
with/IN extraordinary/JJ qualities/NNS to/TO succeed/VB in/IN Mexico/NNP ,/,
"/" says/VBZ Kimihide/NNP Takimura/NNP ,/, president/NN of/IN
Mitsui/NNS*/NNP group/NN 's/POS Kensetsu/NNP Engineering/NNP Inc./NNP
unit/NN ./.

Figure 2
Sample tagged text—**after** correction.

Image from Marcus, M. P., Santorini, B., & Marcinkiewicz M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

Annotated data in linguistics: Textual

- Corpora often come with some kind of annotation
  - Necessary for formal/experimental linguistic analysis
  - Not always provided, though
- Textual annotation can take many forms
  - Sometimes theory dependent
- In NLP and comp ling, raw data is often fine

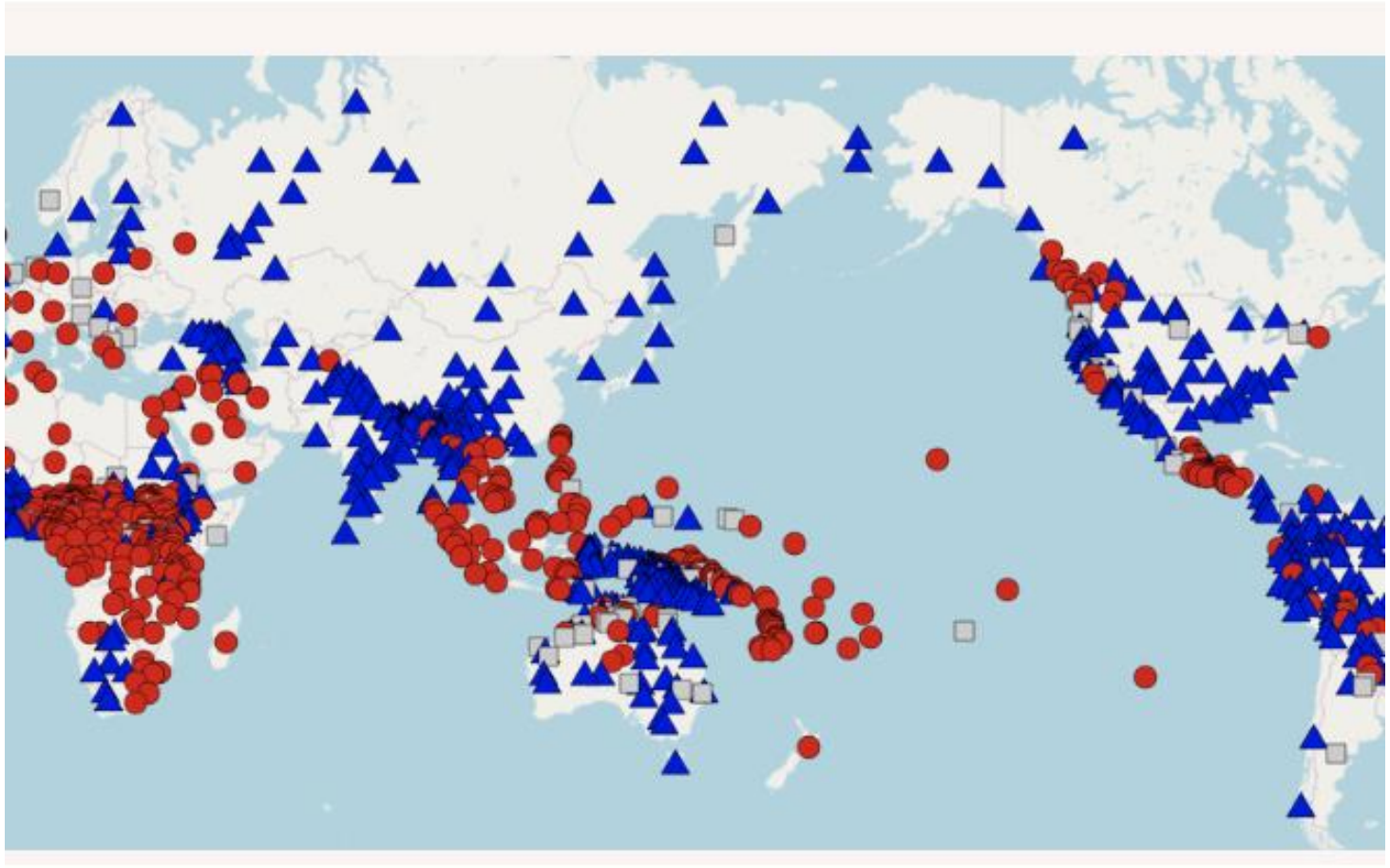


Annotated data in linguistics: Speech

- Annotation is often crucial for analyzing speech data
- Speech is naturally far less structured than text
  - Can't "search" very well in speech data
- Annotations usually label regions or moments in a recording

Linguistics and data science

- The types of corpora we use in linguistics lend themselves to general data science tasks very well
  - Experimental data also often makes use of data science methods
- Examples:
  - Corpus linguistics – a discipline (or maybe just set of methods) that calculate statistics over large amounts of data
  - Sociolinguistics – e.g., finding significant correlations between sociolinguistic variables
  - Phonetics – using thousands of participants to rate the intelligibility of spoken sentences



wals.info/feature/86A?v1=t00d&v3=sccc#2/21.0/152.9 (Dryer, 2005. WALS. Order of Genitive and N

Lx and DS cont.

- Virtually any part of linguistics can involve data science methods
  - Data just needs to be managed in some way
- Not all parts of linguistics make use of these methods though
  - Usually restricted to computational and experimental approaches

Class activity

- Go to the World Atlas of Language Structures (WALS) or PHOIBLE
  - Find one pattern about the world's languages that surprises you, interests you, or contradicts something you assumed
- WALS: good for grammar or word order, e.g.
  - Do most languages put adjectives before or after nouns?
  - How common is SOV vs. SVO word order?
- PHOIBLE
  - Which languages have click sounds?
  - How common are front rounded vowels like /y/?

Corpora in NLP and data science

- All NLP systems are trained on corpora
- Most systems are also tested on corpora
  - Need to be cleaned, maintained, updated, preprocessed, etc.
  - For some data sets, this continual maintenance process never ends (see Davies's English corpora)
- The IMDB data set we are using is a corpus
  - In what ways is it annotated?

Linguistics in NLP and data science

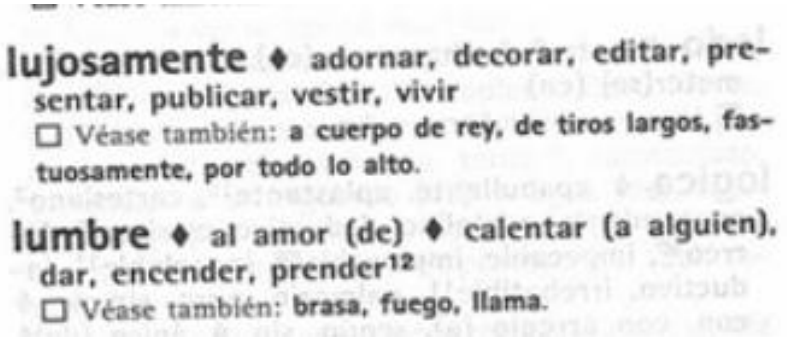
- Consider the following examples
  - *The dog bit the man*
  - *The man bit the dog*
- How do we get different vectors out of these sentences?
  - Some NLP techniques don't consider word order, so these would result in the same vector representation
- Syntactic information can be added to embeddings
  - Out of scope for course, but open area of research

Corpora and data science for linguistics

- NLP uses corpora to engineer new systems
  - Create chat agents
  - Classify different kinds of text along certain criteria
- In linguistics, corpora are used to study language itself
  - Find systematic patterns in language use
  - Used to estimate how often phenomena occur (we've done this already)

Some basic corpus terms

- **Collocate**: a word that appears next to another word
  - Bigrams are examples of collocates (e.g., both *other* and *final* are collocates of the word *the*, as in *the other* and *the final*)
- **Concordance line**: a line of text from a corpus, usually showing linguistic context for a word
- **Phraseme** (sometimes phraseology): a set or fixed phrase that appears in a corpus, e.g., *start a family*



lujosamente ♦ adornar, decorar, editar, presentar, publicar, vestir, vivir
☐ Véase también: a cuerpo de rey, de tiros largos, fastuosamente, por todo lo alto.

lumbre ♦ al amor (de) ♦ calentar (a alguien), dar, encender, prender[12]
☐ Véase también: brasa, fuego, llama.

Types of combination

The dictionary covers the following types of word combination:

Noun entries:

adjective + noun: *bright/harsh/intense/strong* ***light***
quantifier + noun (... of): *a beam/ray of* ***light***
verb + noun: *cast/emit/give/provide/shed* ***light***
noun + verb: ***light*** *gleams/glows/shines*
noun + noun: *a* ***light*** *source*
preposition + noun: *by the* ***light*** *of the moon*
noun + preposition: *the* ***light*** *from the window*

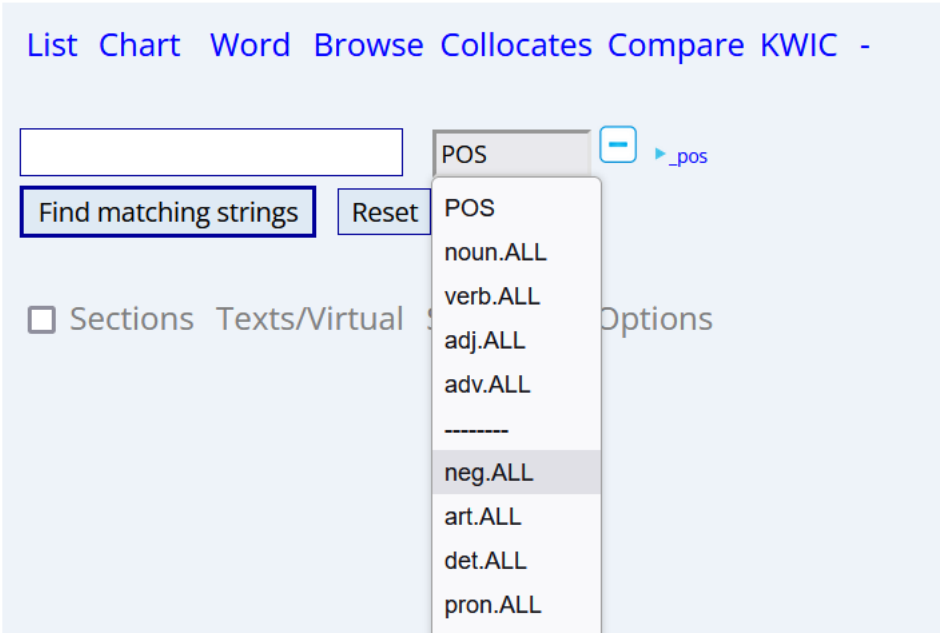
Tangible products from linguistic corpora

- There are collocation/combinatorial dictionaries that will say what words commonly modify other words
- Often aimed toward language learners
  - But can be useful for anyone who writes

Using linguistic corpora

Web interfaces

- When using corpora for linguistic studies a web interface is often used
- Typically use some simple form of RegEx with some application specific codes
- Generally easy to use, but possibly limited



Concordancer programs

- There are some programs that let you use your own custom-built corpora
  - Or corpora you've downloaded
- AntConc is a free one
- Some cost money (like ParaConc)
- Web-based kits like sketch-engine are becoming more common

Custom analysis: Scripts/programs

- For more sophisticated analysis, you just write a script
- Historically, Perl or Bash+Awk were used frequently
  - Pascal sometimes too
- Now, Python is the default scripting tool
  - But you can use whatever you want, really
- Tools like NLTK or spaCy are useful for scripting in Python
  - [spe͡ɪ si] / spay-SEE

Corpus activity

- Go to https://www.english-corpora.org and choose a corpus that seems interesting
- Think about some phrases you are interested in (very new phrases may not appear)
- Play with the search tools to see if you can find any interesting patterns
  - For example, see if you can determine what adjectives typically precede the word *breeze* and what verbs usually occur after *dog*
  - That is, find collocates of *breeze* and *dog*
- The limit is about 20 searches per 24 hours unfortunately :/

Corpora in action I: Aijmer (2021)

"THAT'S WELL GOOD"

(1) S0502: I've got a *real well good* one and I'm *well happy* with mine
S0498: >> I thought your first one I thought your first thought would have been –ANONnameF
S0432: that was second (.)
S0502: I've got a *well good* one (S7KD)[2]

Study overview

- Example of "small-scale" data science
  - With appropriate expert knowledge, not as much data is needed as for systems like neural networks
- Trying to determine how the word *well* intensifies the semantics of particular classes of words
  - Looking at phrases like *well good*, meaning *very good*

Study corpus: British National Corpus

- Uses the British National Corpus 2014
  - 10 million spoken words from British English from 2010s
  - Modeled after BNC from 1994 (which was 10 times larger and had more genres)
  - Can compare with older BNC 1994 corpus
- Aijmer focused on the spoken genre of the corpus
  - Usually transcripts of television shows and news segments

Table 1. Frequency of *Well* Followed by Lexical Adjective in BNC1994D and BNC2014S (Pmw)

	N	Pmw
BNC1994D	53	11.07
BNC 2014S	69	13.76

Some results (pmw = per million words)

Table 2. Frequency of *Well* Followed by Adjectives in BNC1994D

	BNC1994D	Men	Word count	Rate pmw	Women	Word count	Rate pmw
0-14	11 (20.75%)	7	247,560	28.3	4	187,726	21.3
15-24	21 (39.62%)	14	212,977	65.7	7	383,136	18.3
25-34	3 (5.66%)	2	287,983	6.9	1	528,041	1.9
35-44	4 (7.55%)	1	317,356	3.2	3	508,501	5.9
45-59	1 (1.89%)	—	321,379	—	1	538,357	1.9
60+	1 (1.89%)	—	303,508	—	1	480,096	2.1
Unknown	12 (22.64%)	—	—	—	—	—	—
Total	53 (100%)						

Table 3. Frequency of *Well* Followed by Adjectives in the Spoken BNC2014S

	BNC2014S	Men	Word count	Rate pmw	Women	Word count	Rate pmw
0-14	1 (1.45%)	—	21,175	47.2	1	48,179	20.8
15-24	26 (37.68%)	12	339,636	35.3	14	622,710	22.5
25-34	18 (26.09%)	4	186,544	21.4	14	198,677	70.5
35-44	3 (4.35%)	2	251,268	7.8	1	410,949	2.4
45-59	2 (2.90%)	1	197,450	5.1	1	334,578	3.0
60+	—	—	596,346	—	—	343,657	—
Unknown	19 (27.54%)		—	—	—	—	—
Total	69						

More results

Interpretation

- Intensifier *well* used more often among younger speakers
- Mixed results about whether men or women use it more often
- Over time, it seems that it spread regionally in Britain, and from men to women
- One potential explanation afforded for the age patterns was that younger speakers are using it to appear fashionable
  - As many colloquial terms and phrases are used

Corpora in action II: Stange (2021)

"HE SHOULD SO BE IN JAIL"

(1) We are *so* going to get in trouble for this. (SOAP, *AMC*, 2008)[3]
(2) Matthew, trying to get your parents back together is *so* kindergarten. (SOAP, *OLTL*, 2004)
(3) It's so easy to say, and it's *so* impossible to do. (SOAP, *AMC*, 2004)
(4) Well, if it's a party, I'm *so* there. (SOAP, *OLTL*, 2011)
(5) Ambition is *so* yesterday. (SOAP, *OLTL*, 2008)
(6) Fantastic. She is *so* behind us, Kevin. (SOAP, *YR*, 2005)
(7) I'm totally mortified. I was *so* out of line. (SOAP, *GH*, 2008)
(8) Oh, my God, Billy Abbott is *so* not my cup of tea. (SOAP, *YR*, 2008)

Study overview

- Another "small-scale" data science study
- Looking at the use of "preverbal *so*"
  - Pattern originally attributed to Generation X
  - For reference, Greatest Generation -> Silent Generation -> Baby Boomers -> **Generation X** -> Millennials -> Generation Z -> Generation Alpha
  - Generation X individuals *tend* to be parents of Generation Z

(28) Oh Katherine, I *so* appreciate your help. (SOAP, *YR*, 2008)
(29) I do *so* wish that you hadn't come alone, darn it. (SOAP, *ATWT*, 2006)
(30) You're *so* making that up. (SOAP, *OLTL*, 2005)
(31) I *so* would give anything to have my husband come home. (SOAP, *PASS*, 2004)
(32) I am going to *so* fire that hippy-dippy quack. (SOAP, *AMC*, 2003)
(33) Well, I would *so* have to agree with you. (SOAP, *OLTL*, 2007)

Study corpus: The Corpus of American Soap Operas

- 100 million words transcribed from 10 different shows
- Years 2001-2012
- Found 1357 relevant occurrences of preverbal *so*

Table 2. Speaker Age and Gender: Frequency of Preverbal *So* (SOAP)

		<40	40+	?	Total
Tokens	Women	620	414	11	1045
	Men	139	163	6	308
	Total	759	577	17	1353
Characters	Women	181	91	11	283
	Men	75	68	7	150
	Total	256	159	18	433
Tokens/character	Women	3.43	4.55		
	Men	1.85	2.40		

Results I: Sentence types and speaker demographics

- Preverbal *so* tends to occur in declarative sentences
  - Not questions
- Tended to be produced by younger women

	Affirmative uses				Negated uses		
Rank	Verb	*N*	%	Rank	Verb	*N*	%
1	LOOK FORWARD	125	11	1	WANT	33	15
2	WANT	103	9	2	HAPPEN	19	9
3	HOPE	75	7	3	GET	16	7
4	LOVE	60	5	4	GO	9	4
5	APPRECIATE	56	5	5	DESERVE	8	4
6	WISH	54	4	6	BELIEVE	7	3
7	GET	35	3	7	HAVE	6	3
8	ENJOY	32	3	8	KNOW	6	3
9	REGRET	25	2	9	NEED	5	2
10	DESERVE	22	2	10	WORK	5	2
	Total	592	52		Total	114	52

Results II: Types of verbs that can be modified

- In affirmative cases, tends to involve words expressing desire or emotion
- Only 3 words frequent in both affirmative and negated contexts: *want*, *get*, and *deserve*

Interpretation

- Preverbal *so* has lexical (i.e., collocational) restrictions on its use
- There are sociolinguistic patterns in its use, at least in the data analyzed
- This is another example of how using expert knowledge can make it less necessary to use machine learning techniques