

Conducting Corpus-based Research on Spoken Language

Undergraduate Scholars Programme 2021





About the Project

- Learn how to transcribe speech professionally
- Use spoken corpus data for a linguistic research question

About the Corpus



The Student-Transcribed Corpus of Spoken American English

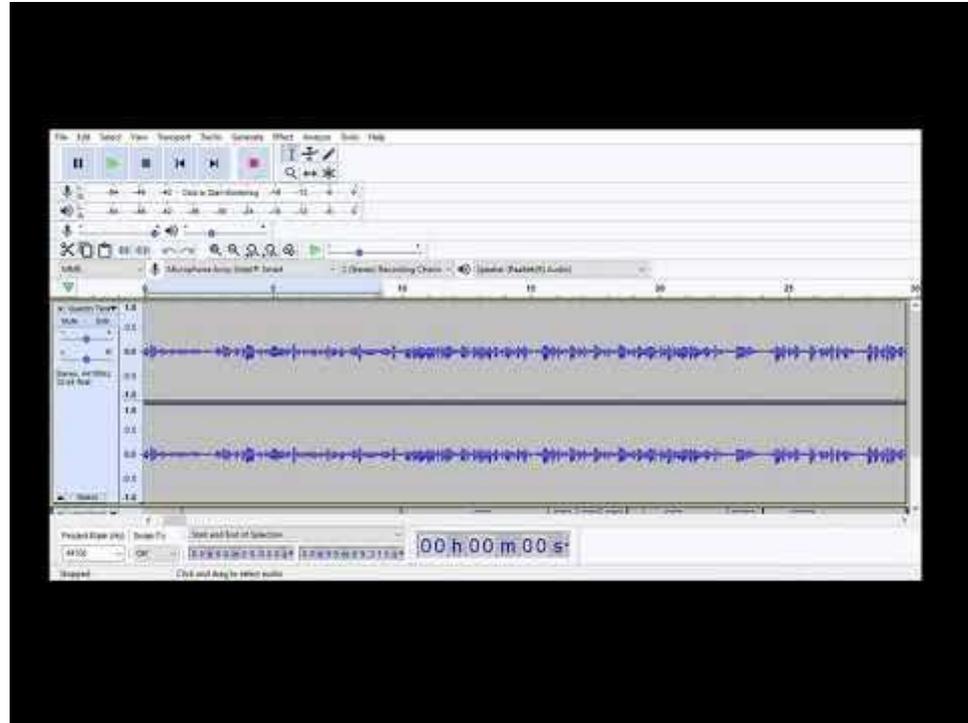
www.SpokenCorpus.org

- Focus on Spoken American English
- Native speaker of American English
- A total of 97,605 word tokens (including disfluencies and punctuation)

Transcription



www.spokencorpus.org



Our Research Question



What are determinants of the variation between the hedges 'sort of' and 'kind of' in spoken American English?

- (1) So that was **sort of** a good experience
(Student-Transcribed Corpus, DCecilHouCar,
Speaker: from Texas, born 1973, education middle)

- (2) Two thousand eighteen was **kind of** a depressing year.
(Student-Transcribed Corpus, MyBoatIsGone,
Speaker: from Florida, born 1984, education middle)

Linguistic hedge: a word or phrase used to express a lack of certainty, commitment, or accuracy, of an utterance.

Our Research Method



1 | 0:00 / 0:03 | I just ER I kind of want to share with you a good enough story line that I do it really have of myself

2 | 0:00 / 0:06 | I just ER I kind of do a lot of free-writing on my personal time that I do it really have of myself

3 | 0:00 / 0:08 | But I just technically just in down, write out some character names, planning kind of them, come up with a story, and then look small on the screen

4 | 0:00 / 0:00 | Results for query: **sort of**

5 | 0:00 / 0:00 | [Back to search](#)

6 | 0:00 / 0:00 | There are 54 hits. Displaying hits 1 to 54

Hit	Audio	Left	Search	Right	File name	Transcriber	Place name	Year of Education
1	0:00 / 0:10	By, by drawing the analogy to the stock sort of I'm	Let's just look at data. Forget our ideologies. What do the data say?	TumDownHeat	anonymous student	Utica	1964.5_VeryHigh	
2	0:00 / 0:02	So that was sort of a good experience		DcecilHouCar	anonymous student	Amarillo	1973.3_Middle	
3	0:00 / 0:04	So that's just a very different sort of approaching ER	way of	DcecilHouCar	anonymous student	Amarillo	1973.3_Middle	
4	0:00 / 0:13	ER, like, a series is just a film drawn out over time, you know, sort of a longer format, particularly this one because it was made by film makers	ER, like, a series is just a film drawn out over time, you know, sort of a longer format, particularly this one because it was made by film makers	DcecilHouCar	anonymous student	Amarillo	1973.3_Middle	
5	0:00 / 0:13	ER you know, they're sort of ER, like a series is just a sort of one because it was made by film makers	ER you know, they're sort of ER, like a series is just a sort of one because it was made by film makers	DcecilHouCar	anonymous student	Amarillo	1973.3_Middle	
6	0:00 / 0:02	So we're hoping to sort of that the next few months	move forward to that the next few months	DcecilHouCar	anonymous student	Amarillo	1973.3_Middle	
7	0:00 / 0:06	ER but I'm... right now, I'm just loving spending time with	ER taking in Atlanta	DcecilHouCar	anonymous student	Amarillo	1973.3_Middle	

There are 54 hits. Displaying hits 1 to 54

1 | 0:00 / 0:10 | By, by drawing the analogy to the stock sort of I'm

2 | 0:00 / 0:02 | So that was sort of a good experience

3 | 0:00 / 0:04 | So that's just a very different sort of approaching ER

4 | 0:00 / 0:13 | ER, like, a series is just a film drawn out over time, you know, sort of a longer format, particularly this one because it was made by film makers

5 | 0:00 / 0:13 | ER you know, they're sort of ER, like a series is just a sort of one because it was made by film makers

6 | 0:00 / 0:02 | So we're hoping to sort of that the next few months

7 | 0:00 / 0:06 | ER but I'm... right now, I'm just loving spending time with

And ER I felt, you know, there was a kind of public purpose in ... in doing that.

```
R Console
> data = read.table(file.choose(), header=TRUE, sep=";", fill=TRUE, row.names=ID)
> sort(data)
data.frame: 161 obs. of 10 variables:
 $ lit      : int  46 71 72 78 9 31 32 33 34 35 ...
 $ Hedge    : chr  "Mead" "Mead" "Mead" "Mead" ...
 $ YearBirth : int  1964 1964 1964 1964 1969 1969 1969 1969 1969 1969 ...
 $ AgeRecording : int  29 29 29 29 40 30 30 30 30 ...
 $ Latitude   : num  30.3 30.3 30.3 30.3 40.0 ...
 $ Longitude  : num  -97.7 -97.7 -97.7 -97.7 -74 ...
 $ Education  : chr  "M_Sci" "M_Sci" "M_Sci" "M_Sci" ...
 $ SpeakerName : chr  "Madeline Miehler [YogaWithMadriene]" "Madriene Miehler [YogaWithMadriene]" ...
 $ Transcriber : chr  "Madriene Miehler [YogaWithMadriene]" "Madriene Miehler [YogaWithMadriene]" ...
 $ Example    : chr  "But we're gonna start inosporulating those sort of things" ...

> head(data)
   lit  Hedge YearBirth AgeRecording Latitude Longitude Education
1  46  Mead    1964         29  30.27  -97.74  2_Low
2  71  Mead    1964         29  30.27  -97.74  2_Low
3  72  Mead    1964         29  30.27  -97.74  2_Low
4  73  Mead    1964         29  30.27  -97.74  2_Low
5  9  Mead    1964         29  40.78  -74.02  2_Low
6  31  Mead    1969         30  43.27  -97.20  2_MidHigh

1  Madriene Miehler [YogaWithMadriene]  Transcriber
2  Madriene Miehler [YogaWithMadriene]  Wenjing Shi

# cross-tabulate speaker, age and use of the hedge hedge
change <- crosstabs(speaker, Age ~ Hedge)
# look at data
change

# install another package called "scales" for colours
install.packages("scales")
library(scales)

# Make a plot
plot(change$YearBirth, change$Kind, change$Kind~change$Ageors,
# plot titles here
```

Effect of age on speakers' use of

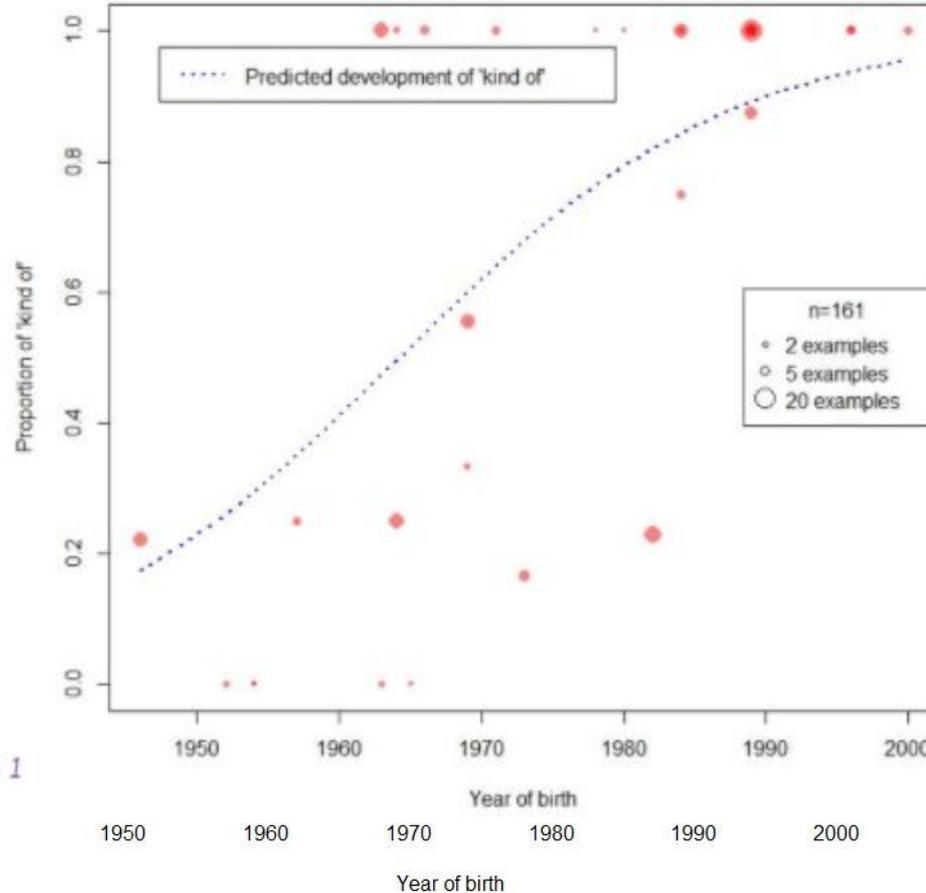


Figure 1

Figure 1

Effect of dialectal area on speakers' use of "kind of" and "sort of"

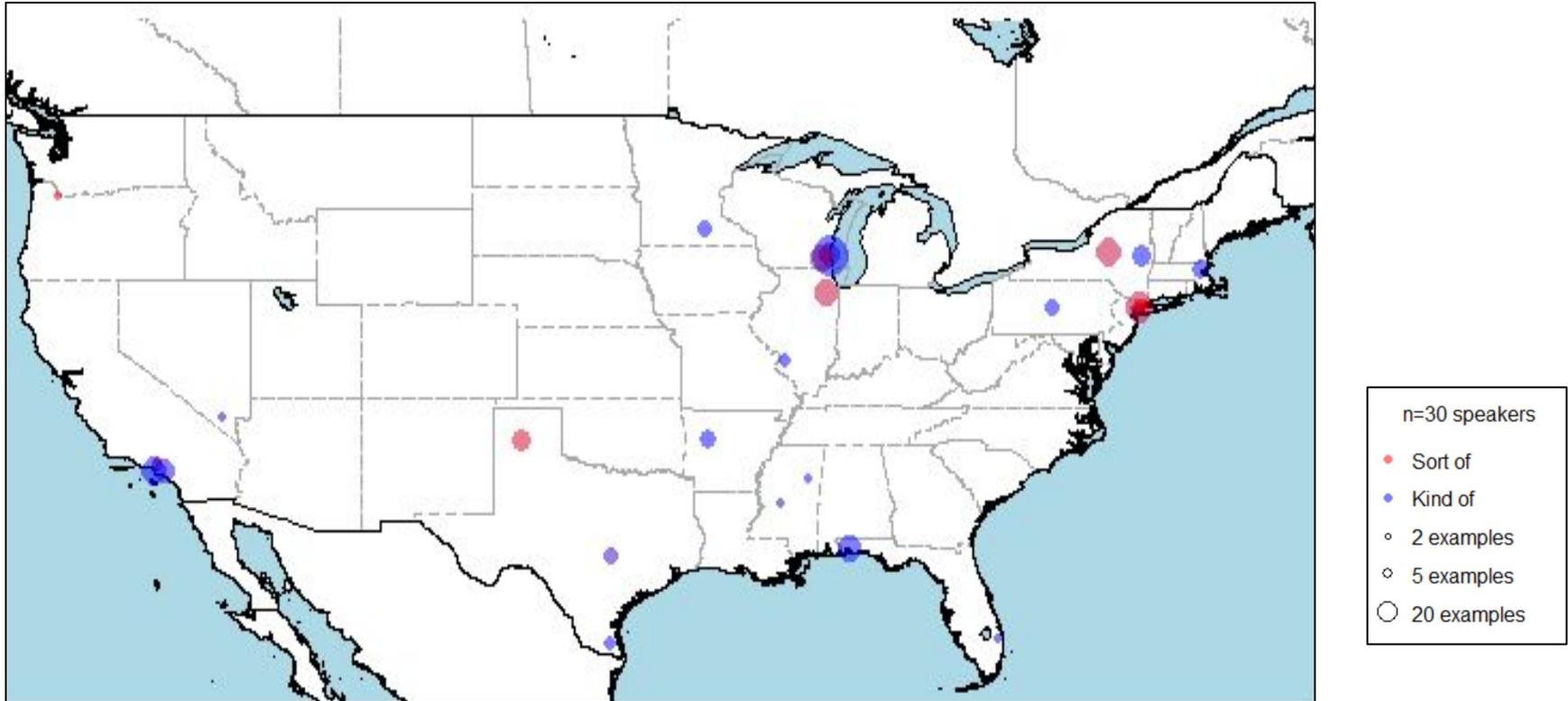


Figure 2

Effect of education on speakers' use of "kind of" and "sort of"



($\chi^2=15.1$, $df=1$, $p<0.001^{***}$)

(odds ratio: 0.1, 95%CI: 0.02-0.36)

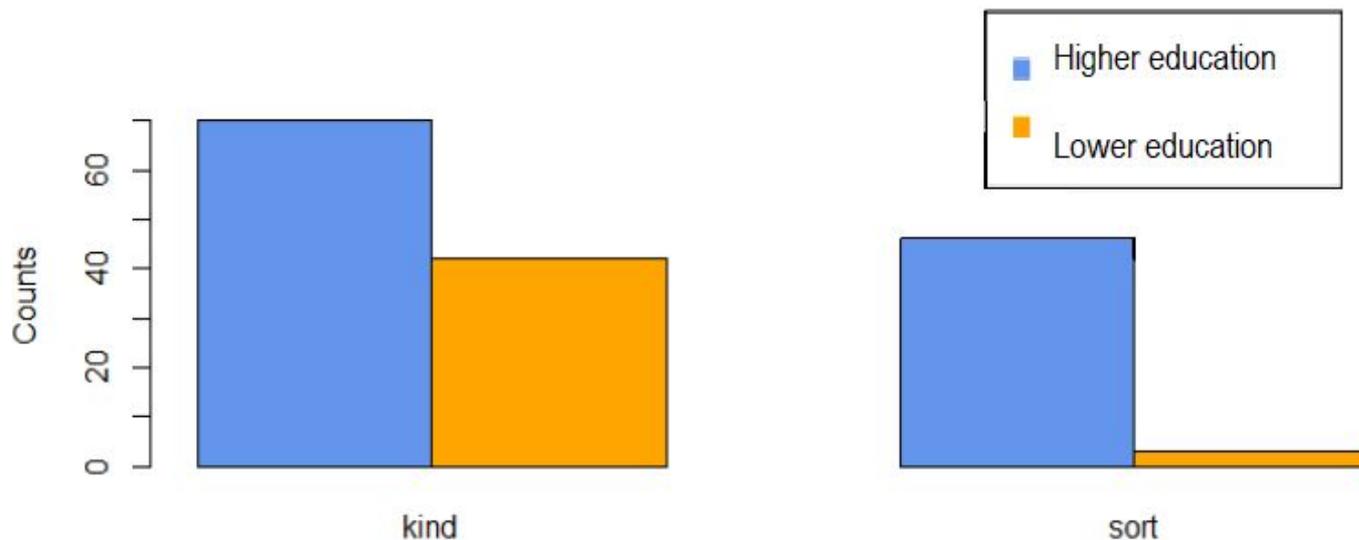


Figure 3

What We Observed



“sort of”

- More conservative
- Possibly used more in northern states
- Mostly used by people with a high education

“kind of”

- Becoming more popular
- Dispersed more inconsistently
- Used by people of both lower education and higher education, but more by lower educated people

Conclusion



- Our findings are still not too clear because the corpus is small
- We would be interested in revisiting our research question again in the future when the corpus is larger
- Useful skills we've learned:
 - How to do professional transcription
 - Using audio editing software
 - Using R to analyse data