German and English Multi-Lingualism

Lillian Carlson
lkc43@pitt.edu
4/21/25

Table of contents

01 Research Guidance

Goals and Questions

Research Questions + Goals

Questions:

- Are Bilingual and Monolingual speakers identifiable?
- What features are more common in which groups of speakers?
- What features are more common in spoken or written language productions and how does the language of production change this?

Goals:

- Find if these differences are identifiable
- Find what features are identifiable

Main Focus:

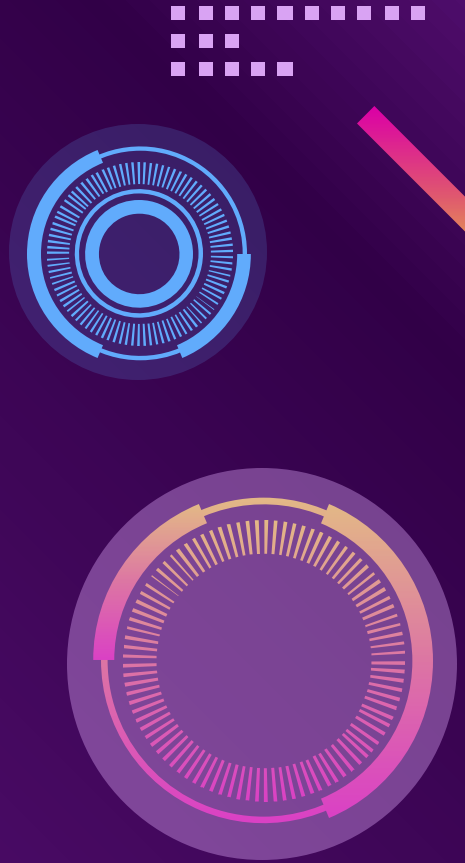
- POS usage
- Two languages to compare cross-linguistic differences of bilingual differences verse monolingual speakers

Related Article Findings

There are several papers written on my corpus, and some look at differences between bilingual and monolingual speakers

- Explicitness
  - Found that Heritage speakers use a greater amount of explicitness in *some* areas/situations
- Discourse and Pragmatic Functions
  - Did find specific differences when looking at multi-use lexical words in terms of elaborations
- V-Final in Russian
  - Russian and German languages contact have begun to shape one another in Heritage Speakers

All of these are interesting but I to look specifically at POS usage and order, as that is pretty different cross-linguistically

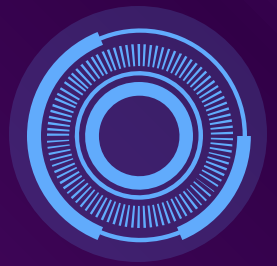


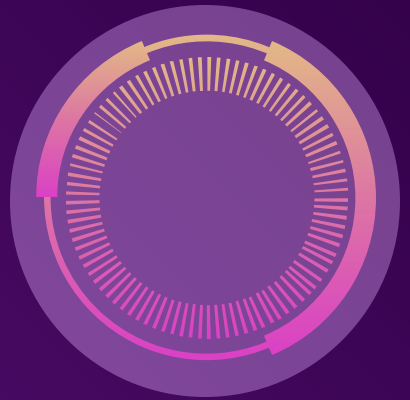
02 Corpus

Research Unit Emerging Grammars (RUEG) corpus
Humboldt University of Berlin

RUEG Corpus Contents

- ConLL-U Files
- Annis
- Audio Files
- ExMaralada
- Pepper workflows?
- PDFs (documentation, Thesis)
- 10.7 GB!! Zipped!!
- 0.3.0 (4/24/20) vs 1.0.0 (4/30/24)
    - Newest Version all together is much larger, but the zipped audio files are separate (thankfully)

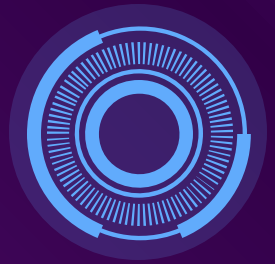


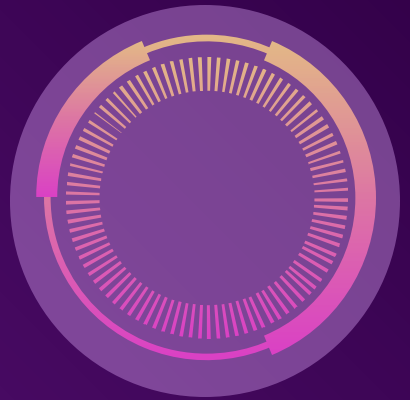


RUEG Corpus

RUEG Corpus is a corpus that contains 5 languages (Greek, Turkish, Russian, English, and German) with lots of variables

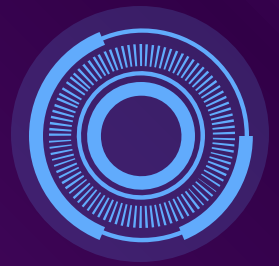
- Formality
- Spoken/Written
- Age Group (adolescent/adult)
- Country of elicitation
- Mono/Bi-Lingual
- Gender
- Many more inside the metadata files all anonymized

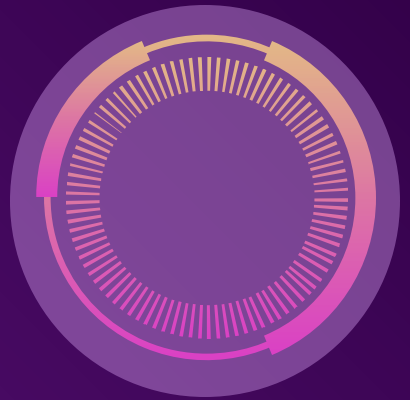


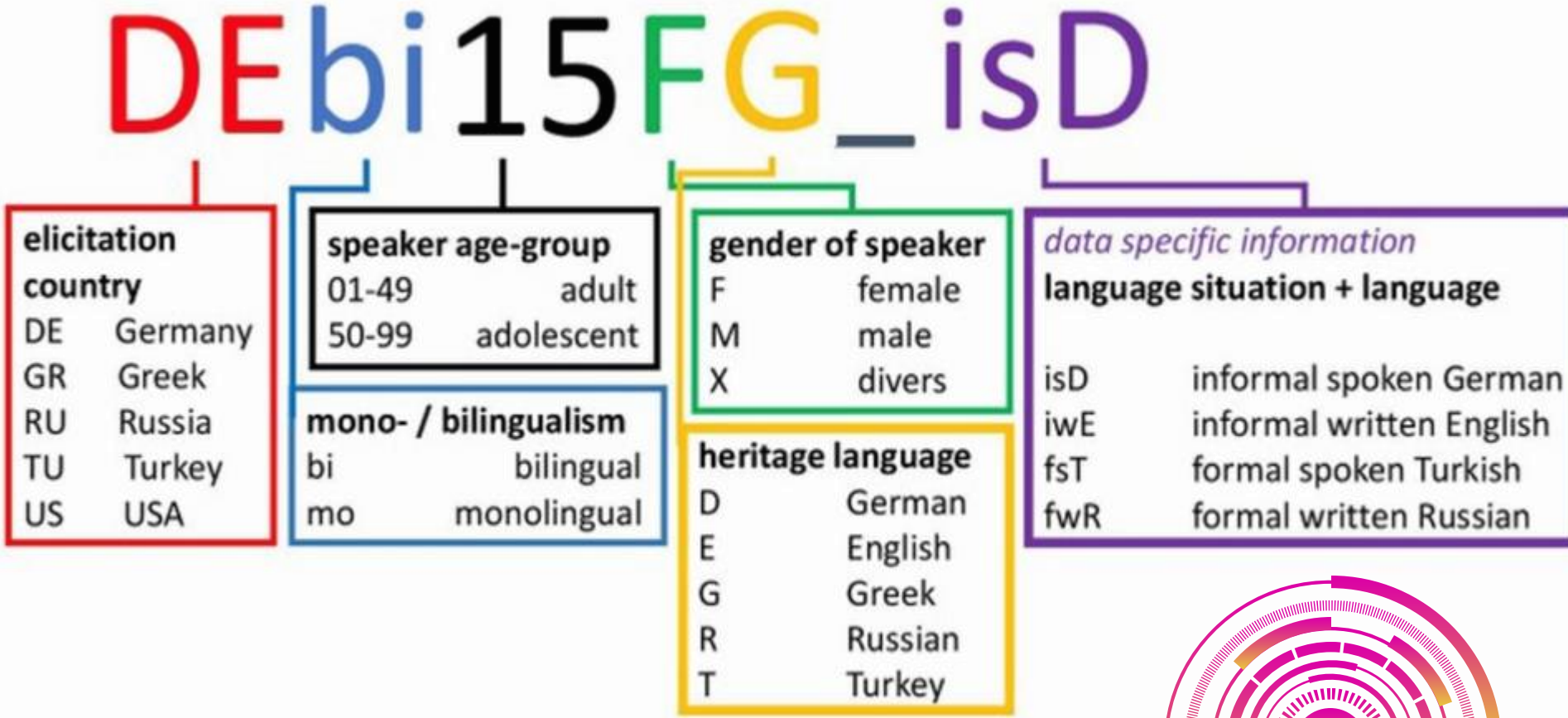


RUEG Corpus

- Speakers are mainly German and American
- Heritage languages (multiple languages spoken at home)
- Fake police reports to get the more natural speech/writing
    - Inflates certain content words
- Almost all automatic
    - Exceptions: speech transcriptions and corrections






DEbi15FG_isD
elicitation country
DE Germany
GR Greek
RU Russia
TU Turkey
US USA
speaker age-group
01-49 adult
50-99 adolescent
mono- / bilingualism
bi bilingual
mo monolingual
gender of speaker
F female
M male
X divers
heritage language
D German
E English
G Greek
R Russian
T Turkey
data specific information
language situation + language
isD informal spoken German
iwE informal written English
fsT formal spoken Turkish
fwR formal written Russian

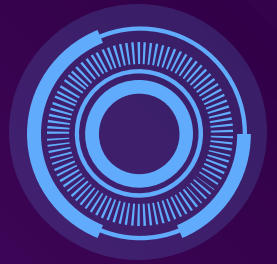
03

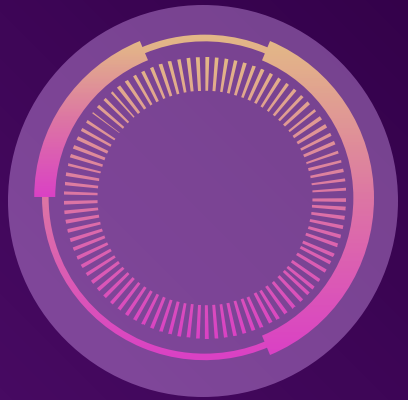
Processing

Manual, SpaCy, and Stanza

Failed Experiments

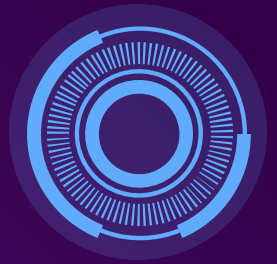
- Manual Parsing
  - Mostly just to play with the ConLL format and understand how ConLL processing works and what ConLL is
- SpaCy Parsing
  - SpaCy was VERY particular and struggled with parsing everything if it wasn't perfect
  - Did not work for me

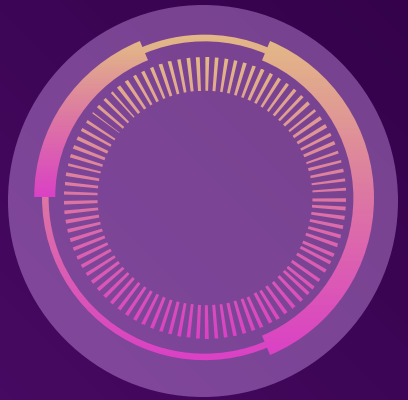




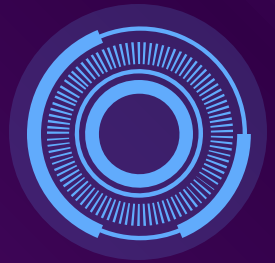
Stanza

- Cons
  - Very limited ConLL supported
    - Could only work things into a dictionary, which I had already done with my manual attempts
- Pros
  - Actually processed all my data

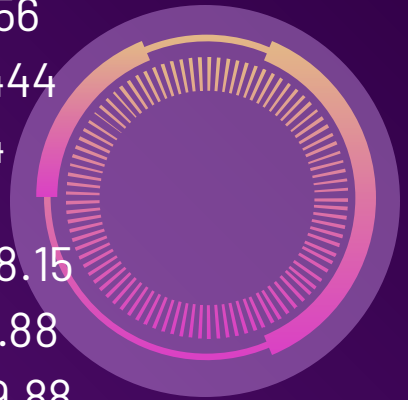




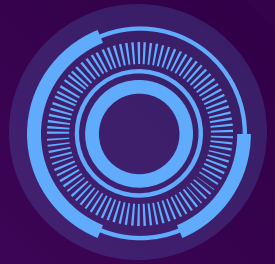
Extraction
Text and POS Lists



- Unigrams
  - Multi-Lingual German- 4773
  - Monolingual German- 1761
  - Multi-Lingual English- 4385
  - Monolingual English- 621
- Bigrams
  - Multi-Lingual German- 4187
  - Monolingual German- 1505
  - Multi-Lingual English- 3941
  - Monolingual English- 557
- Sentences
  - Multi-Lingual German- 586
  - Monolingual German- 256
  - Multi-Lingual English- 444
  - Monolingual English- 64
- Average Words per Sentence
  - Multi-Lingual German~ 8.15
  - Monolingual German~ 6.88
  - Multi-Lingual English~ 9.88
  - Monolingual English~ 9.70



Stanza
Metadata



- Also collected metadata into a dataframe
- Currently has gone unused
    - Future exploration
- What is crucial are the heritage language numbers

```
Index: 1306 entries, DEmo17MD_fsD to USbi04FD_fsE
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Filename           1306 non-null   object
 1   Mono/Bilingual     1306 non-null   object
 2   Language_of_Data   1306 non-null   object
 3   Mode               1306 non-null   object
 4   Formality          1306 non-null   object
 5   Gender             1306 non-null   object
 6   Heritage_Language  1306 non-null   object
 7   Age_Group          1306 non-null   object
 8   Country_of_Data    1306 non-null   object
```

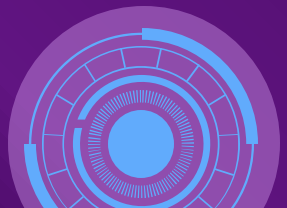
Stanza
Metadata- Heritage Languages

- Only looking at English and German elicitations, but within my data there are different heritage languages
- Mostly Indo-European except Turkish which may prove to give interesting results down the line
    - Turkish is an SOV language, agglutinative, Turkic family
    - Greek is pro-drop with pretty free word order

```
There are 327 German heritage language data files
There are 64 English heritage language data files
There are 260 Turkish heritage language data files
There are 267 Greek heritage language data files
There are 388 Russian heritage language data files
```

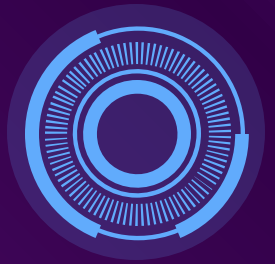
04

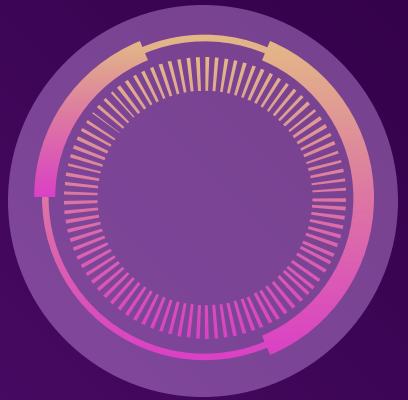
Analysis



With the Human Eye

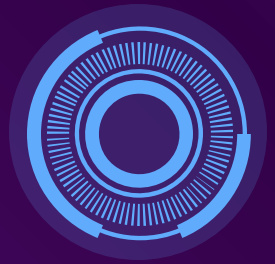
- ”What I found most interesting was the fact that the driver of the first car checked to see if he hit anything in the front of his car because assessing the damage to the back of his car.”
- “He then went to check on the owners of the vehicle to see if they were okay”

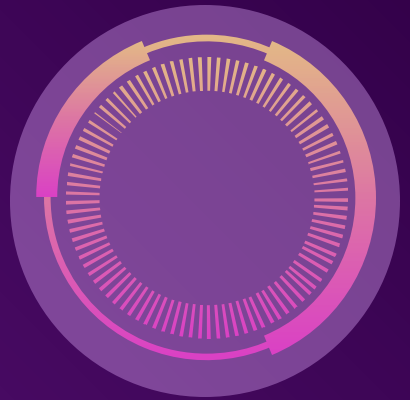




POS Analysis

Unigram- Overview





- Raw counts are hard to compare (text size)
- Top categories all similar
- More adpositions within English while there are more adverbs within German

```
[('NOUN', 702), ('DET', 618), ('VERB', 558), ('ADV', 550), ('PRON', 448), ('AUX', 354)
[('NOUN', 251), ('DET', 210), ('ADV', 207), ('VERB', 189), ('PRON', 176), ('AUX', 127)
[('NOUN', 685), ('VERB', 675), ('DET', 673), ('ADP', 420), ('PROPN', 330), ('CCONJ', 2
[('DET', 102), ('VERB', 99), ('NOUN', 92), ('ADP', 61), ('CCONJ', 44), ('PROPN', 43),
```

POS Analysis
Unigram- Text

- Looks like the top German word is ‘and’- weird that it’s not ‘the’?
- No clear difference here between Mono/Multi-lingual speakers

```
[(('und', 'CCONJ'), 221), (('.', 'PUNCT'), 175), (('die', 'DET'), 155), (('der', 'DET'),
[(('und', 'CCONJ'), 76), (('.', 'PUNCT'), 66), (('die', 'DET'), 37), (('der', 'DET'), 35)
[(('the', 'DET'), 428), (('and', 'CCONJ'), 220), (('.', 'PUNCT'), 158), (('car', 'NOUN'),
[(('the', 'DET'), 62), (('and', 'CCONJ'), 39), (('car', 'NOUN'), 35), (('.', 'PUNCT'), 19
```

POS Analysis

Unigram- POS

- More pronoun usage in the monolingual
- More auxiliary usage in multilingual
- Hard to see a whole flat with flat rates!

```
[('NOUN', 1387),
 ('DET', 1291),
 ('VERB', 1233),
 ('ADV', 784),
 ('ADP', 712),
 ('CCONJ', 609),
 ('AUX', 605),
 ('PRON', 539),
 ('ADJ', 503),
 ('PUNCT', 451),
 ('PROPN', 381),
 ('INTJ', 250),
 ('PART', 142),
 ('SCONJ', 119),
 ('NUM', 58),
 ('SYM', 29)]
```

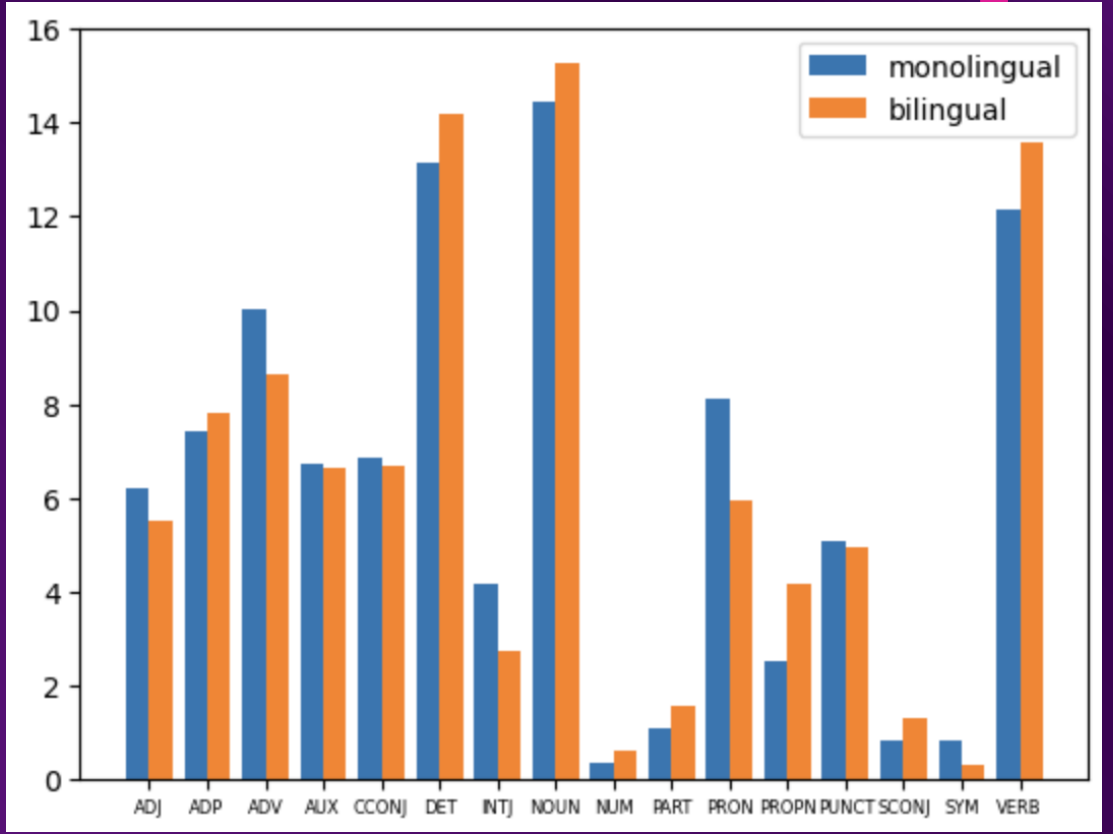
Bilingual

```
[('NOUN', 343),
 ('DET', 312),
 ('VERB', 288),
 ('ADV', 238),
 ('PRON', 193),
 ('ADP', 176),
 ('CCONJ', 163),
 ('AUX', 160),
 ('ADJ', 147),
 ('PUNCT', 121),
 ('INTJ', 99),
 ('PROPN', 60),
 ('PART', 26),
 ('SCONJ', 20),
 ('SYM', 20),
 ('NUM', 9)]
```

Monolingual

POS Analysis
Unigram- POS

- Normalized results!
- Lexical categories are more common in the bilingual data despite greater data amount



POS Analysis
Bigram- Text

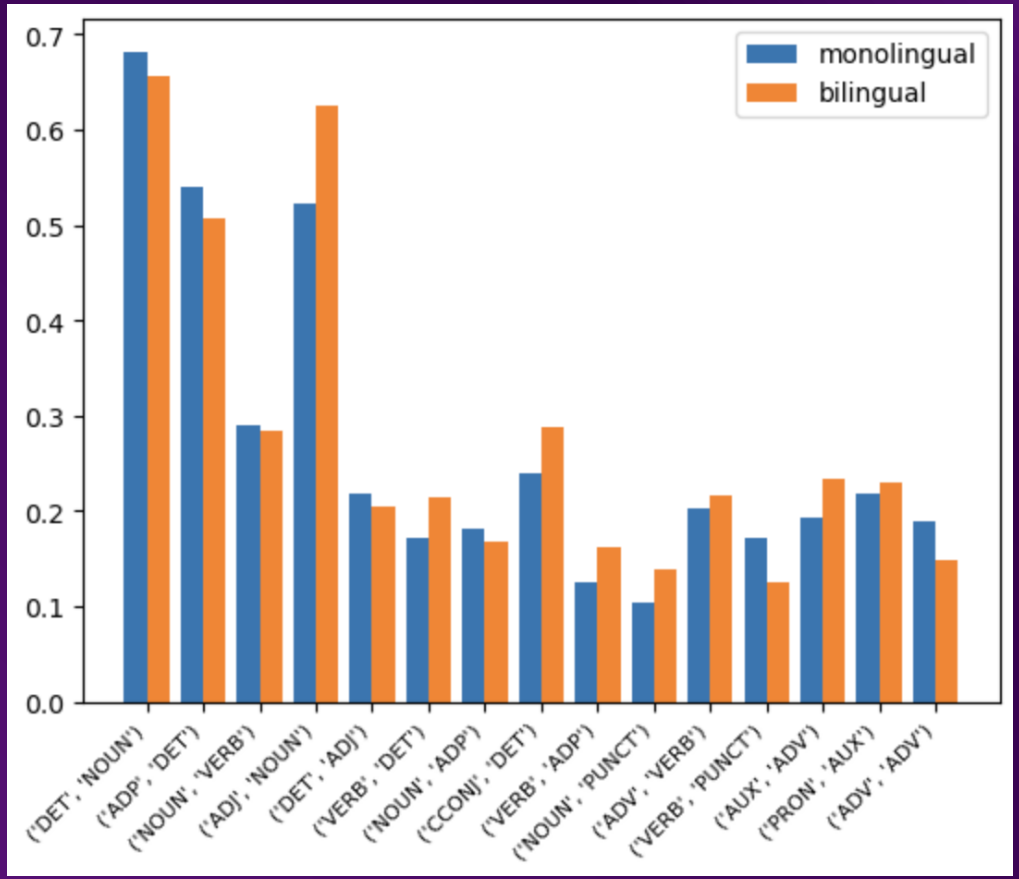
- Shocking how many lexical words
  - prompted speech
- Some obvious ones with POS, but some splits too
- Still very hard to look at

```
[(('DET', 'NOUN'), 439), (('NOUN', 'VERB'), 236), (('ADJ', 'NOUN'), 154), (('ADP', 'DET'), 149), (('DET', 'ADJ'), 131), (('VE
[(('DET', 'NOUN'), 146), (('NOUN', 'VERB'), 70), (('ADP', 'DET'), 68), (('ADJ', 'NOUN'), 54), (('DET', 'ADJ'), 47), (('VERB',
[(('DET', 'NOUN'), 407), (('ADP', 'DET'), 209), (('VERB', 'DET'), 172), (('VERB', 'ADP'), 152), (('NOUN', 'ADP'), 139), (('DE
[(('DET', 'NOUN'), 66), (('VERB', 'ADP'), 26), (('ADP', 'DET'), 26), (('VERB', 'DET'), 25), (('DET', 'ADJ'), 21), (('NOUN', '
```

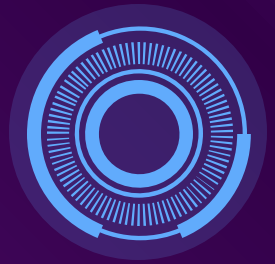
```
[((('die', 'DET'), ('Polizei', 'NOUN')), 108), ((('und', 'CCONJ'), ('der', 'DET')), 31), ((('Polizei', 'NOUN'), ('gerufen', '
[(('die', 'Polizei'), 19), (('mit', 'dem'), 13), (('das', 'war'), 9), (('ja', 'das'), 8), (('das', 'Auto'), 7), (('der', 'Frau
[(('and', 'the'), 49), (('the', 'car'), 43), (('the', 'police'), 42), (('of', 'the'), 39), (('called', '911'), 36), (('and',
[(('the', 'car'), 17), (('the', 'other'), 9), (('and', 'the'), 9), (('of', 'the'), 7), (('car', 'behind'), 7), (('when', 'the
```

POS Analysis
Bigram- Text

- Again, normalized
- Took top 15 most popular bigrams
- Differences between adjective noun bigrams
- Amounts don't matter- what's more important are the differences



POS Analysis
Trigram Overview



- Also went over trigrams
- Nothing too interesting
- Counts are similar despite gap in amount of data

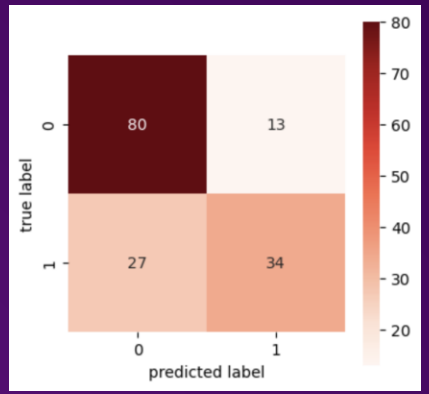
```
[(('die', 'Polizei', 'gerufen'), 7690), (('die', 'Polizei', '.'), 4534), (('die', 'Polizei', 'angerufen'), 4486), (('das', 'wa
[(('mit', 'dem', 'Ball'), 1124), (('Mann', 'mit', 'dem'), 1069), (('die', 'Polizei', 'gerufen'), 698), (('aus', 'den', 'Autos
[(('called', 'the', 'police'), 5828), (('the', 'blue', 'car'), 5637), (('the', 'first', 'car'), 3629), (('the', 'car', 'behin
[(('the', 'other', 'one'), 168), (('when', 'the', 'car'), 146), (('the', 'car', 'behind'), 138), (('car', 'behind', 'it'), 124
```

```
[(('DET', 'NOUN', 'VERB'), 597), (('DET', 'ADJ', 'NOUN'), 463), (('ADP', 'DET', 'NOUN'), 389), (('NOUN', 'VERB', 'PUNCT'), 258
[(('ADP', 'DET', 'NOUN'), 648), (('DET', 'NOUN', 'VERB'), 499), (('DET', 'ADJ', 'NOUN'), 433), (('NOUN', 'ADP', 'DET'), 281),
[(('ADP', 'DET', 'NOUN'), 601), (('VERB', 'DET', 'NOUN'), 536), (('DET', 'ADJ', 'NOUN'), 504), (('DET', 'NOUN', 'ADP'), 434),
[(('ADP', 'DET', 'NOUN'), 584), (('VERB', 'DET', 'NOUN'), 581), (('DET', 'NOUN', 'ADP'), 578), (('DET', 'ADJ', 'NOUN'), 327),
```

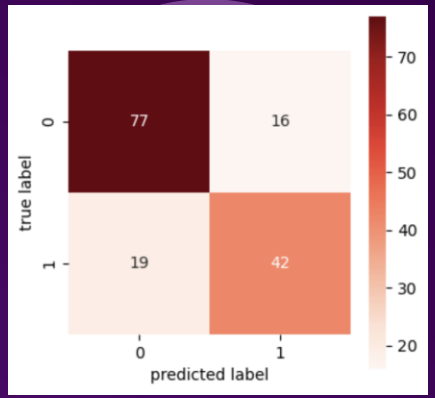
POS Analysis
Machine Learning

- Classifier
  - Feature extraction
- Baseline: 58.39%
- Unigrams: 74. 03%
- Bigrams: 77.3%
- Trigrams: 79.87%

	Text	Tag
153	INTJ	MONO
164	ADJ ADJ PUNCT	MONO
116	CCONJ DET NOUN PUNCT	MONO
513	VERB ADP PROPN ADV	BI
487	CCONJ VERB DET NOUN	BI
433	DET PRON AUX VERB ADV PUNCT	BI
358	VERB PROPN	BI
606	VERB ADP PROPN ADV	BI
495	CCONJ VERB DET NOUN PUNCT	BI
717	SCONJ PRON VERB DET ADJ PRON PART NOUN	MONO



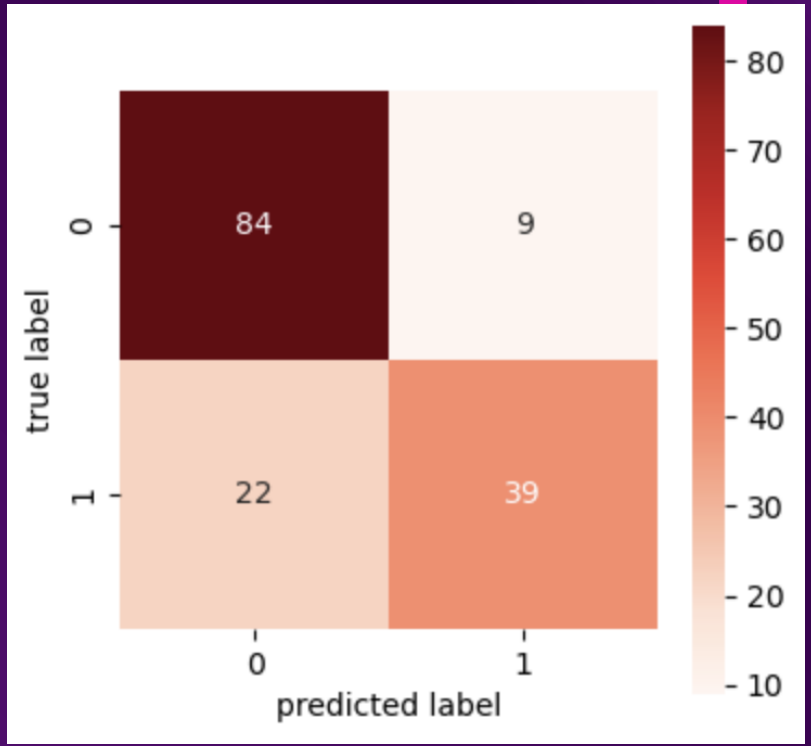
Unigrams



Bigrams

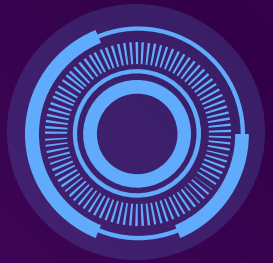
POS Analysis
Machine Learning- Trigrams

- Feature Extraction
- Part of features with the highest log probability values



BI: NOUN, DET, VERB, PROPN, DET_NOUN, ADP, CCONJ, AUX, INTJ, ADV, ADJ, VERB_DET, PUNCT, VERB_PROPN, ADP_DET,
MONO: INTJ, NOUN, DET, PRON, ADV, VERB, DET_NOUN, ADP, SYM, ADJ, PROPN, PUNCT, CCONJ, AUX, NOUN_VERB, ADP_DET

POS Analysis
Machine Learning Testing



- K-Fold Testing with 5 splits

```
[0.8051948051948052,
 0.7857142857142857,
 0.7987012987012987,
 0.7792207792207793,
 0.7843137254901961]
```

Unigram

```
[0.8246753246753247,
 0.7467532467532467,
 0.7922077922077922,
 0.7792207792207793,
 0.7712418300653595]
```

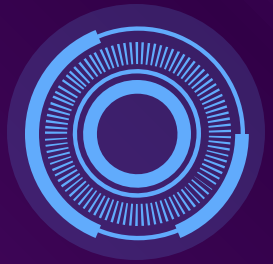
Bigram

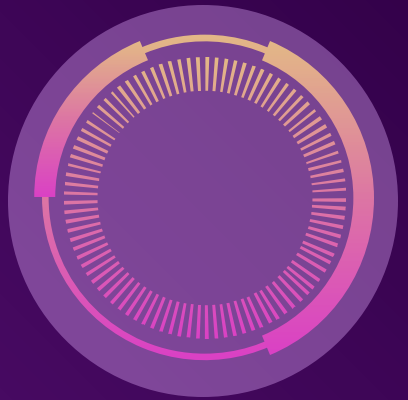
```
[0.7207792207792207,
 0.7012987012987013,
 0.7792207792207793,
 0.7337662337662337,
 0.6862745098039216]
```

Trigram

POS Analysis
Machine Learning Testing

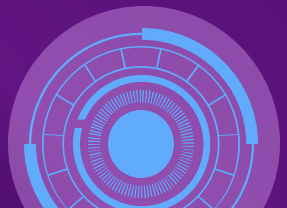
- Despite likely overfitting
  - results still show that there are correlations that humans cannot see between bilingual and monolingual speech





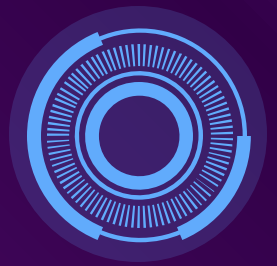
05

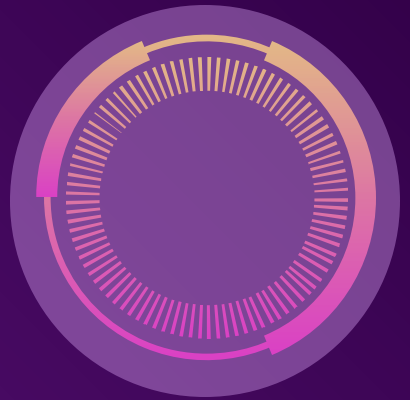
Conclusions



Conclusions

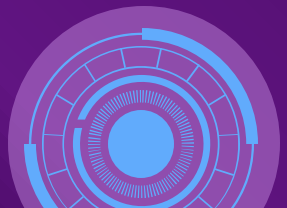
- There are linguistic and syntactic differences between the way monolingual and bilingual speakers speak regardless of language
- Some of these ways are pronoun usage, adverb usage, auxiliary usage and determiner usage
- A lot more work should be done in a to focus specific words or parts of speech
- Dependency relations





06

Future Steps



Near Future Steps

Written vs Spoken Speech

How does this affect accuracy in detecting lingualism?

Topic Modeling

With the use of topic modeling, could a distinction be made between the different heritage languages?

Further Future Steps

Dependency Relation Analyzing

The contexts and relationships that certain words are used may vary depending on lingualism

Speech Analysis

So much speech data! I would love to see if there are any differences in speech patterns/accents depending on lingualism!

Sources

Brehmer, B., & Usanova, I. (2015). Let's fix it? Cross-linguistic influence in word order patterns of Russian heritage speakers in Germany. In H. Peukert (Ed.), *Transfer effects in multilingual language development* (pp. 161–188). https://doi.org/10.1075/hsld.4.08bre. Amsterdam: John Benjamins.

Kellemar. (2024, May 22). *Rueg corpus*. Institut für Linguistik. https://www.linguistik.hu-berlin.de/en/institut-en/professuren-en/rueg/rueg-corpus

Labrenz, A. (2023). Functional Variation of German Also across Registers and Speaker Groups. *Contrastive Pragmatics*, *4*(2), 289-320. https://doi.org/10.1163/26660393-bja10077

Pashkova, T., & Allen, S. (2025). Explicitness of referring expressions in heritage speakers' majority English. *Lingua*, *Volume 314*(103854). https://doi.org/10.1016/j.lingua.2024.103854.

Credits + More

lkc43@pitt.edu
Repo Here: GitHub Repo

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**

Please keep this slide for attribution