Kazakh-Russian Code-Switching Analysis

DS4LING 2340 | May 2, 2023 | Moldir Baidildinova

Outline

- Background of the Project
- Data Sourcing
- Data Annotation
- Analysis
- Conclusion
- Limitations

Background

Linguistic Phenomenon

- **Code-switching:** individuals use two or more languages within a single conversation or even within a single utterance

Target languages

- **Kazakh** – a Turkic language with an agglutinative structure
- **Russian** – a Slavic language with an analytical structure

Socio-Political context

- **Kazakhstan:** a Post-Soviet Union country where Russian is dominant in many aspects of community life

Project Goals

- to carry out an **explanatory analysis** of Kazakh-Russian CS based on the conversational dataset
- to investigate **structural and syntactic types of CS** through **linguistic annotation**
- to examine whether Kazakh-Russian **bilingualism is balanced** and whether the **language shift** is happening toward L2 Russian.

Data Sourcing

- the conversational dataset is sampled from the **IARPA Babel Program Kazakh language collection release** by LDC at UPenn
- contains approximately **203 hours** of Kazakh **conversational and scripted telephone speech** collected in 2013 and 2014 along with corresponding transcripts.
- the Kazakh speech in this release represents that spoken in **the Northeastern and Southern dialect** regions of Kazakhstan.
- The gender distribution among speakers is approximately equal; speakers' ages range from 16 years to 64 years.

Data Annotation

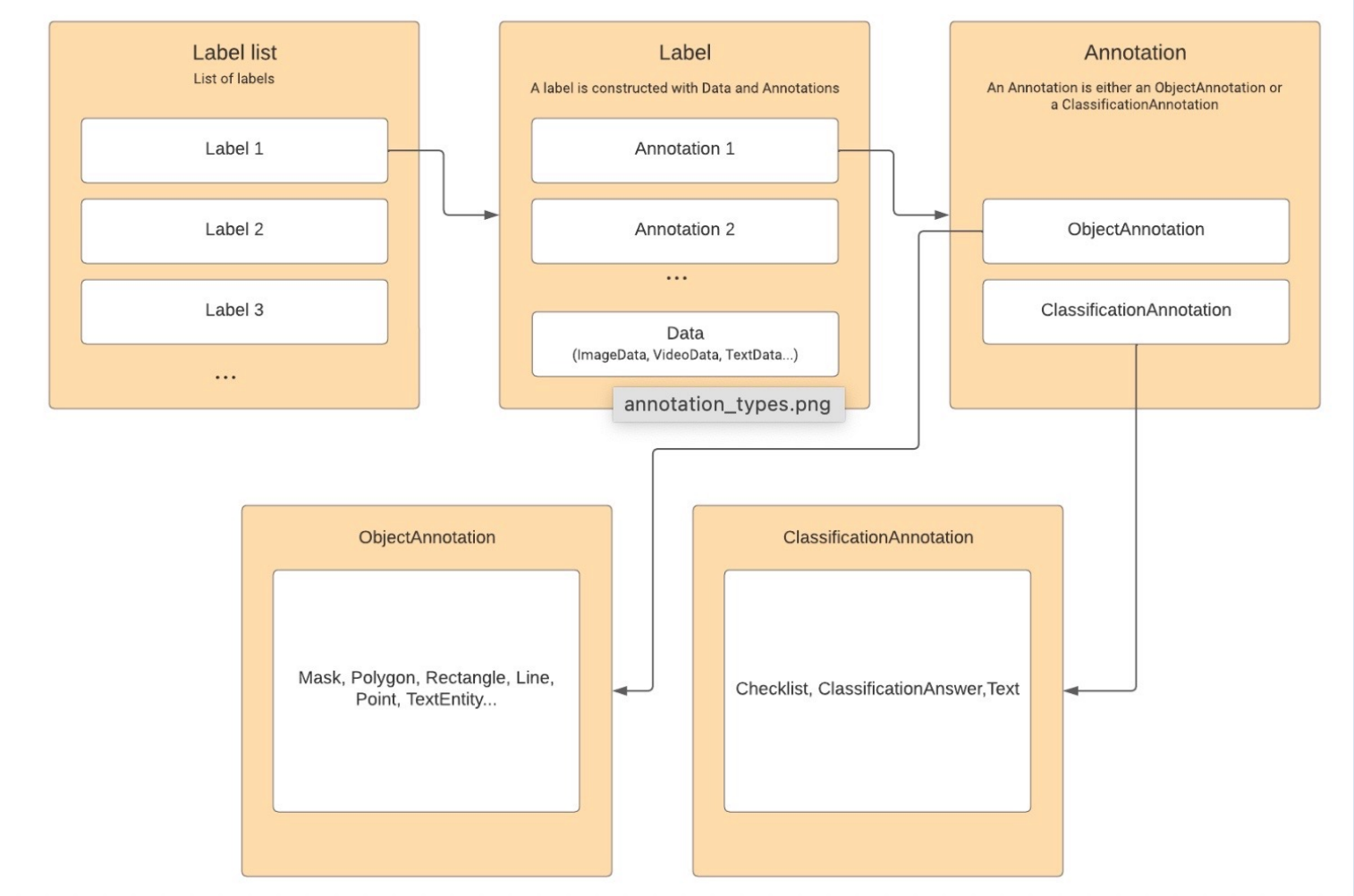
- **An annotation scheme** was developed based on **Dyachkov et al.(2020)** who annotated and analyzed CS in four corpora of minority languages of Russia.

1 tag	Utterance - uttr
kz	Kazakh
rs	Russian

2 tag	Code-switching types - cs
inter-sentential	1st utterance in Kaz and 2nd in Rus or vice-versa
uttr	utterance (since this is a spoken data)
intra-sentential	within one utterance
disc	discourse marker
phr	phrase (a group of words)
vp	verbal phrase, Russian verb+Kazakh verbs+affixes
cl	clause
intra-word	words that have an alternative in Kazakh
n	noun
adj	adjective
adv	adverb
pn	pronoun
conj	conjunction
interj	interjection
morph	Russian stem (any POS) with Kazakh affixes
part	particles
vb	verbs
other	Russian numericals, slang words, etc.

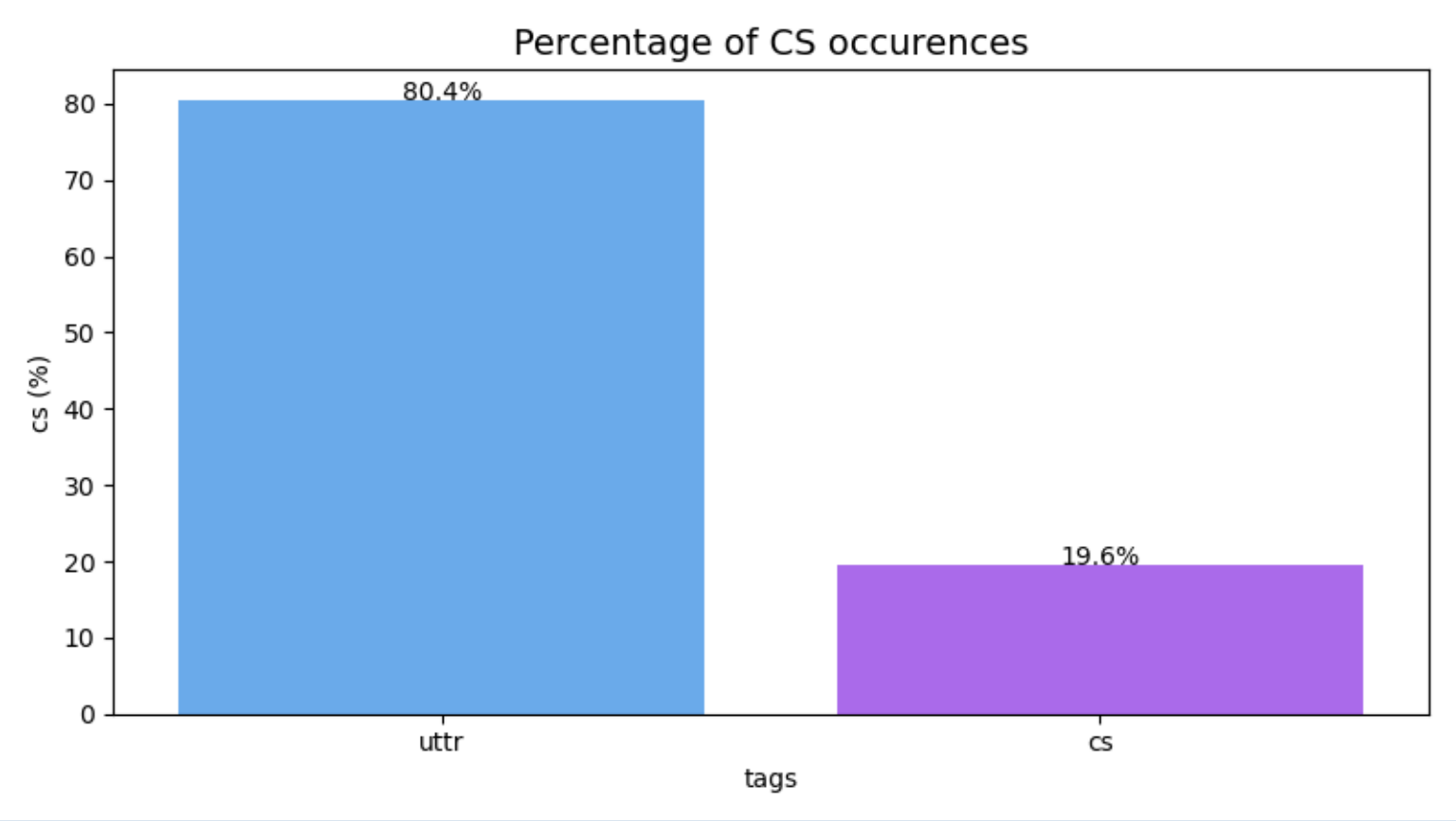
Data Annotation

- Next, fifty (50) text files have been annotated using **the Labelbox platform**
- The Labelbox platform offers **a standoff annotation** (offline) format where annotations are stored separately from the annotated text which can be retrieved via the API key.



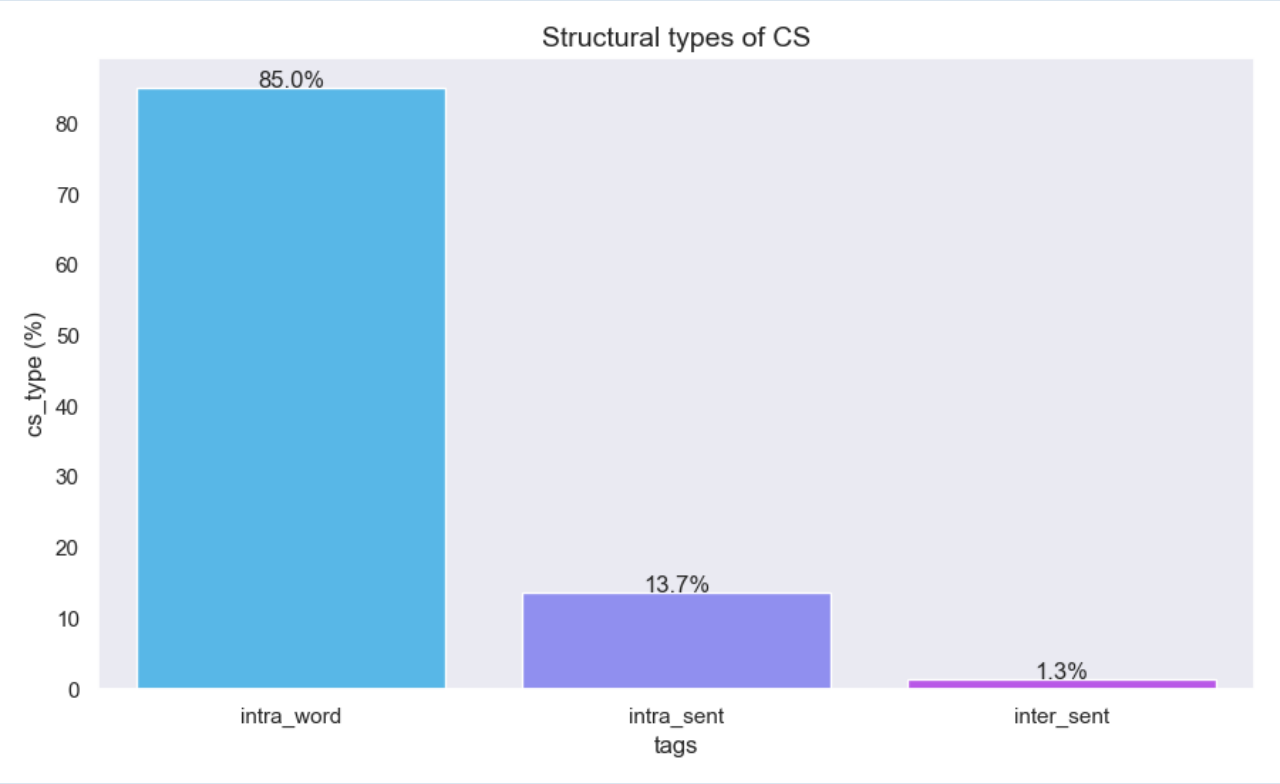
Analysis

- Overall, **3071** utterances were annotated and **601** occurrences of **CS** were observed.



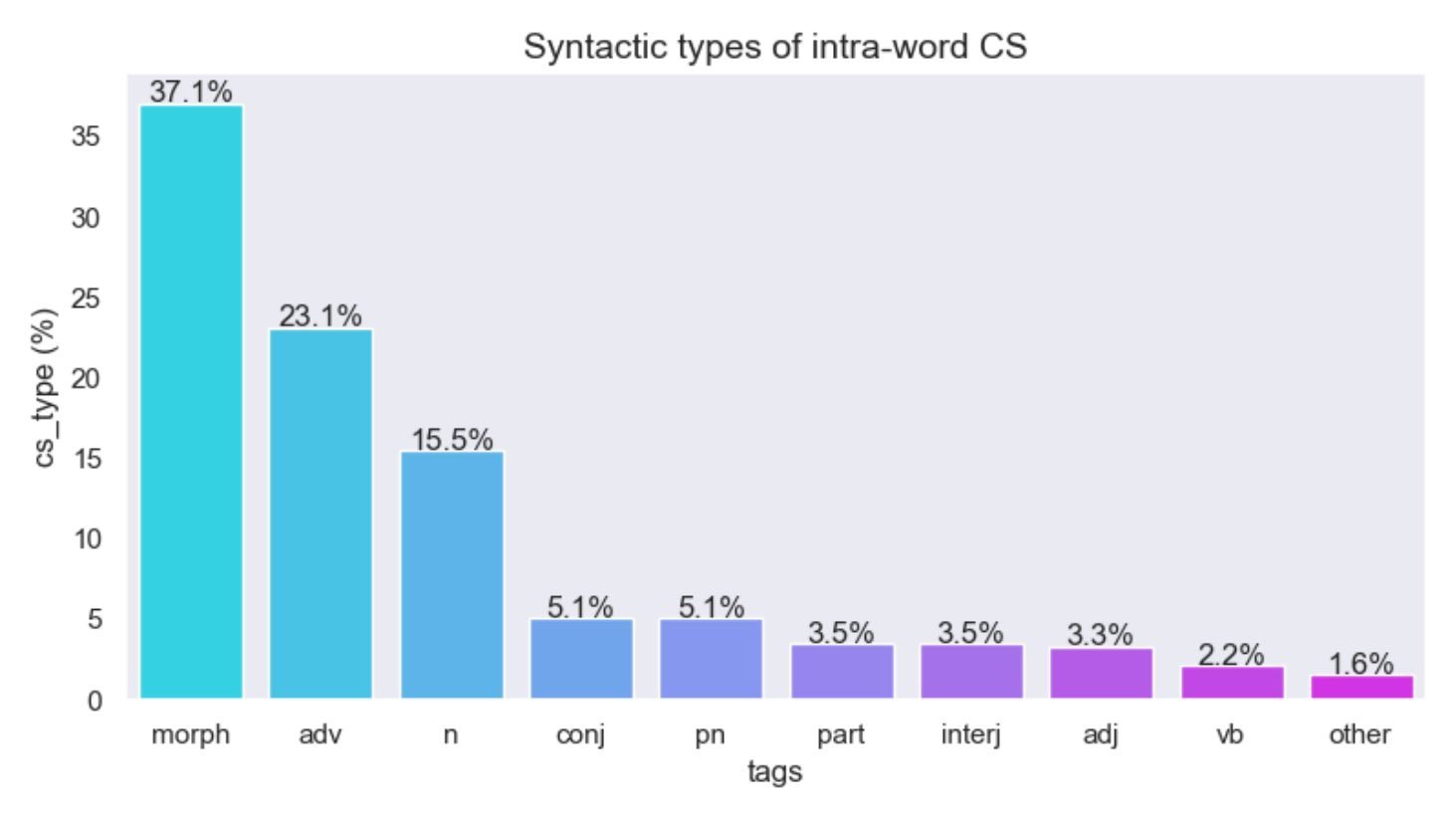
Analysis

- Word-internal shifts are significantly prevalent compared to intra- and inter-sentential CS.



Analysis

- Within the word-internal shifts the most common syntactic constituent is morphemes - Russian stems followed by Kazakh affixes.



Analysis

- Russian stems are highly integrated into the clause structure so they behave like Kazakh constituents.

- *Example 1:*

  > **продавец+ке** бар шығар орын
  > **seller.Rus+DAT.Kaz** [there is] might position
  > *There is might be a position for a seller.*

- *Example 2:*

  > апайдың **настроение+сі+не** байланысты дейсің ғой
  > teacher.GEN **mood.Rus+poss.Kaz+DAT.Kaz** depend say well
  > *Well, it depends on teacher's mood.*

Analysis

- Interestingly, speakers tend to use both Russian and Kazakh variations of the word in one utterance.
- Example 3:

  > **поняла** ұқтың ба

  > **understand.Rus** understand.Kaz part.Kaz

  > *Do (you) understand?*

- Example 4:

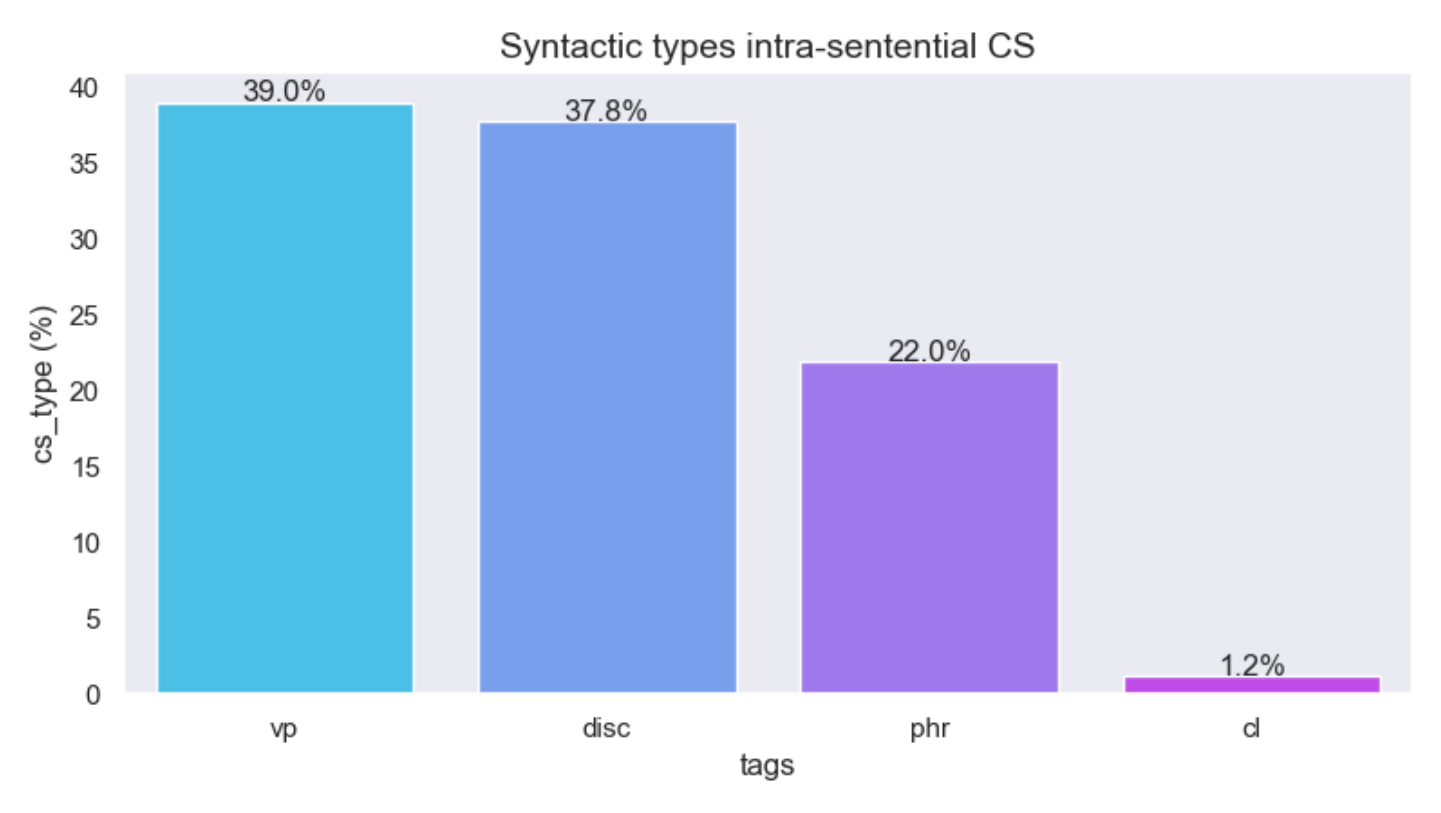
  > барамыз ғой тағы **ещё**

  > come.Fut well again.Kaz **again.Rus**

  > *Well, we will come again.*

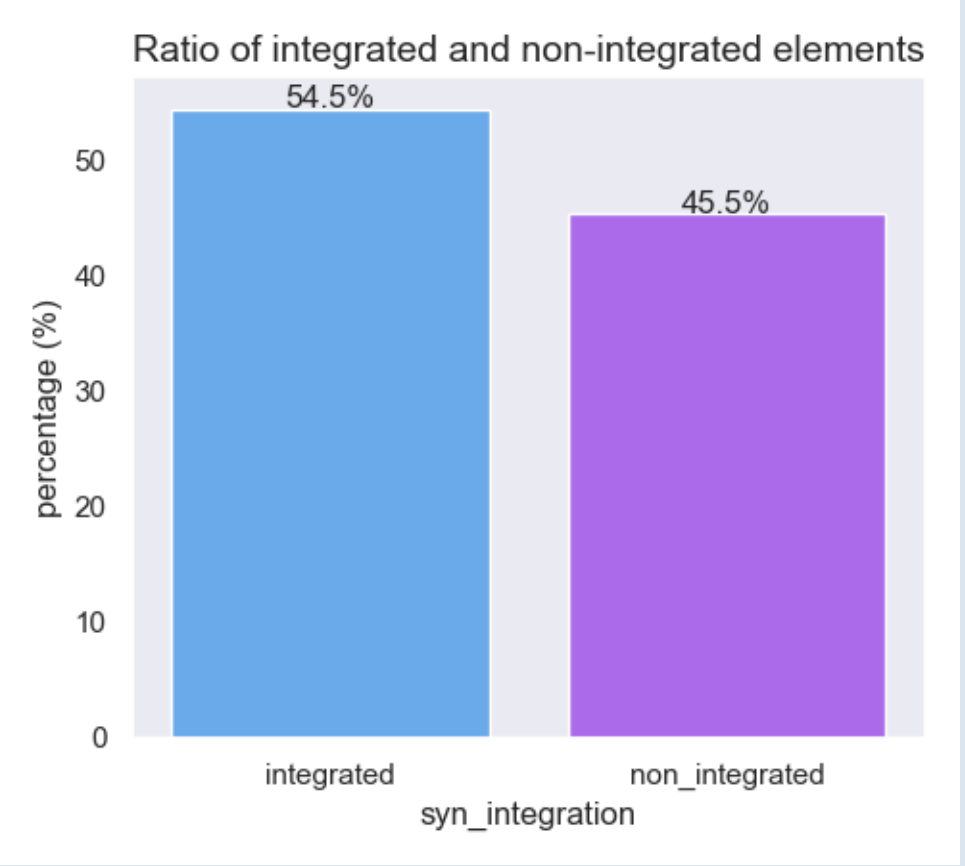
Analysis

- Among intra-sentential CS verbal phrases and discourse markers are also the most common types:



Analysis

- Following Dyachkov et al.(2020) experiment, I calculated the overall ratio of integrated and non-integrated elements by collapsing the structural types of CS.
- For example, elements marked with Kazakh affixes and/or followed by modalities are deemed as **integrated**, while single-word Russian insertions are deemed as **non-integrated**, as shown in the table below.



Analysis

- Another fact that accounts for imbalanced bilingualism is that the speakers frequently use the following mismatched adjectives and nouns:
  - высш**ий** (masculine) образован**ие** (neuter) юридическ**ий** (masculine)
  - военн**ый** (masculine) кафедр**а** (feminine)

- In Russian a preceding adjective should align in gender with a proceeding noun:
  - высш**ее** образован**ие** юридическ**ое** - *all elements are marked with neutral endings*
  - военн**ая** кафедр**а** - *both elements are marked with feminine endings*

Conclusion

Current Project

- Imbalanced bilingualism
  - Structurally word-internal shifts (85%) are prevalent
    - The most common syntactic types are morphemes, adverbials, and nouns
- Early stage of language shift
  - the extremely low percentage of inter-sentential CS (1.3%)
- Russian stems are highly integrated into the clause structure so they behave like Kazakh constituents

Limitations

Current Project

- An annotator bias
- Manual annotation
- Limited annotated samples
- Educational background of speakers

Future Replication

- Two annotators (at least)
- Automated or semi-automated annotation
- Bigger sample size
- Adding speakers with different educational background

Kөp rahmet!

Thank you!

mob75@pitt.edu