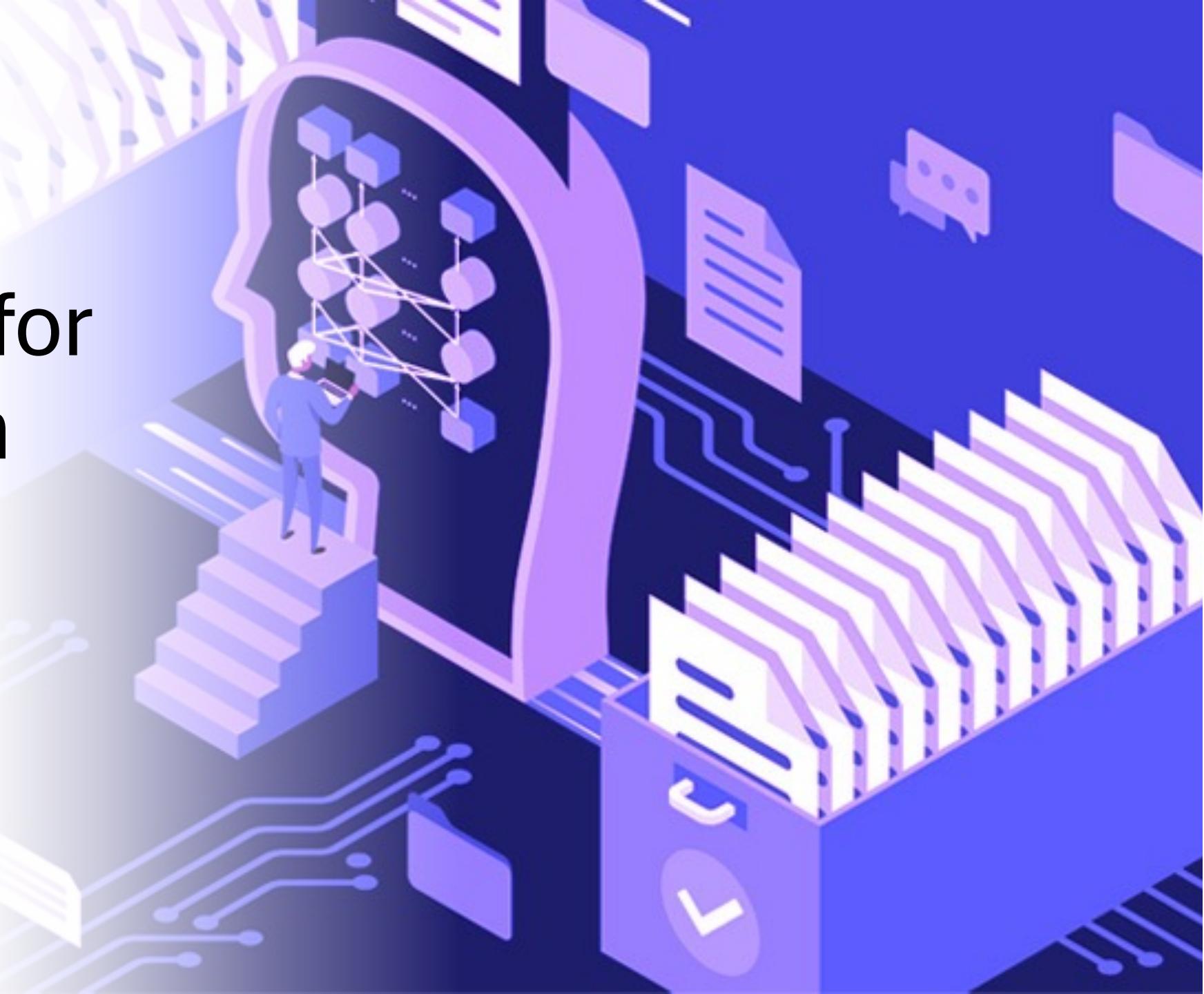


Annotations for *MultiplEYE* in DE, UK, RU



Content

- Data
 - Structure
 - Ukrainian
 - Russian
 - German
- Tokenization and POS Tagging
- Semantic role labeling, Evaluation
- Sentence Alignment
- Next steps

Data structure

*For each language, we have
the same short translated
texts.*

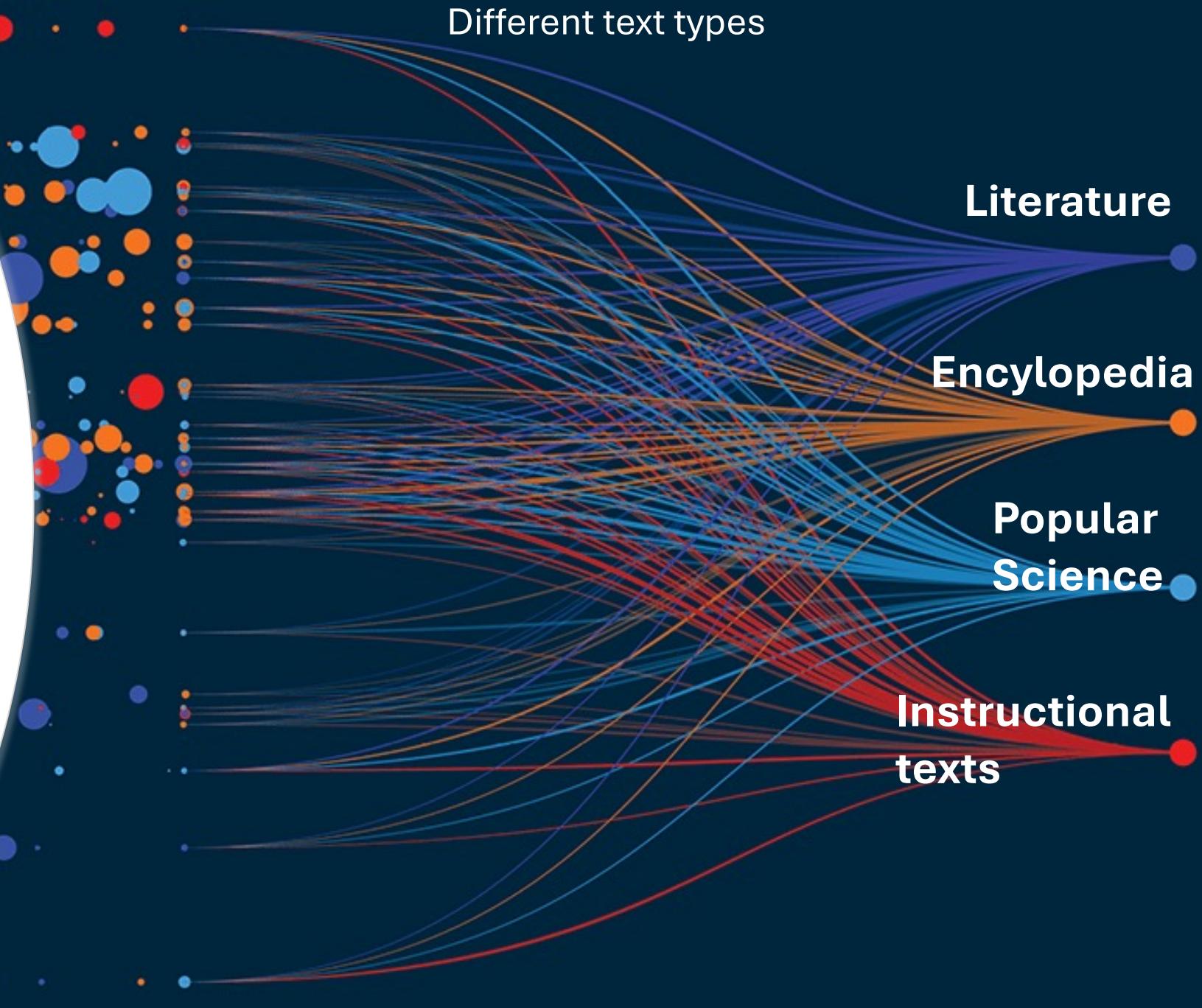
Different text types

Literature

Encyclopedia

Popular
Science

Instructional
texts



Ukrainian data

„Містер і місіс Дурслі, що жили в будинку номер чотири на вуличці Прівіт-драйв, пишалися тим, що були, слава Богу, абсолютно нормальними.“ (Harry Potter)

„Беручи до уваги, що визнання гідності, яка притаманна всім членам людскої сім'ї, і рівних та невід'ємних їх прав є основою свободі, справедливості та загального миру;“ (Human Rights)

Russian data

«Мистер и миссис Дурслы проживали в доме номер четыре по Тисовой улице и всегда с гордостью заявляли, что они, слава богу, абсолютно нормальные люди. Уж от кого-кого, а от них никак нельзя было ожидать, чтобы они попали в какую-нибудь странную или загадочную ситуацию.»
(Harry Potter)

«Принимая во внимание, что признание достоинства, присущего всем членам человеческой семьи, и равных и неотъемлемых прав их является основой свободы, справедливости и всеобщего мира;» (Human Rights)

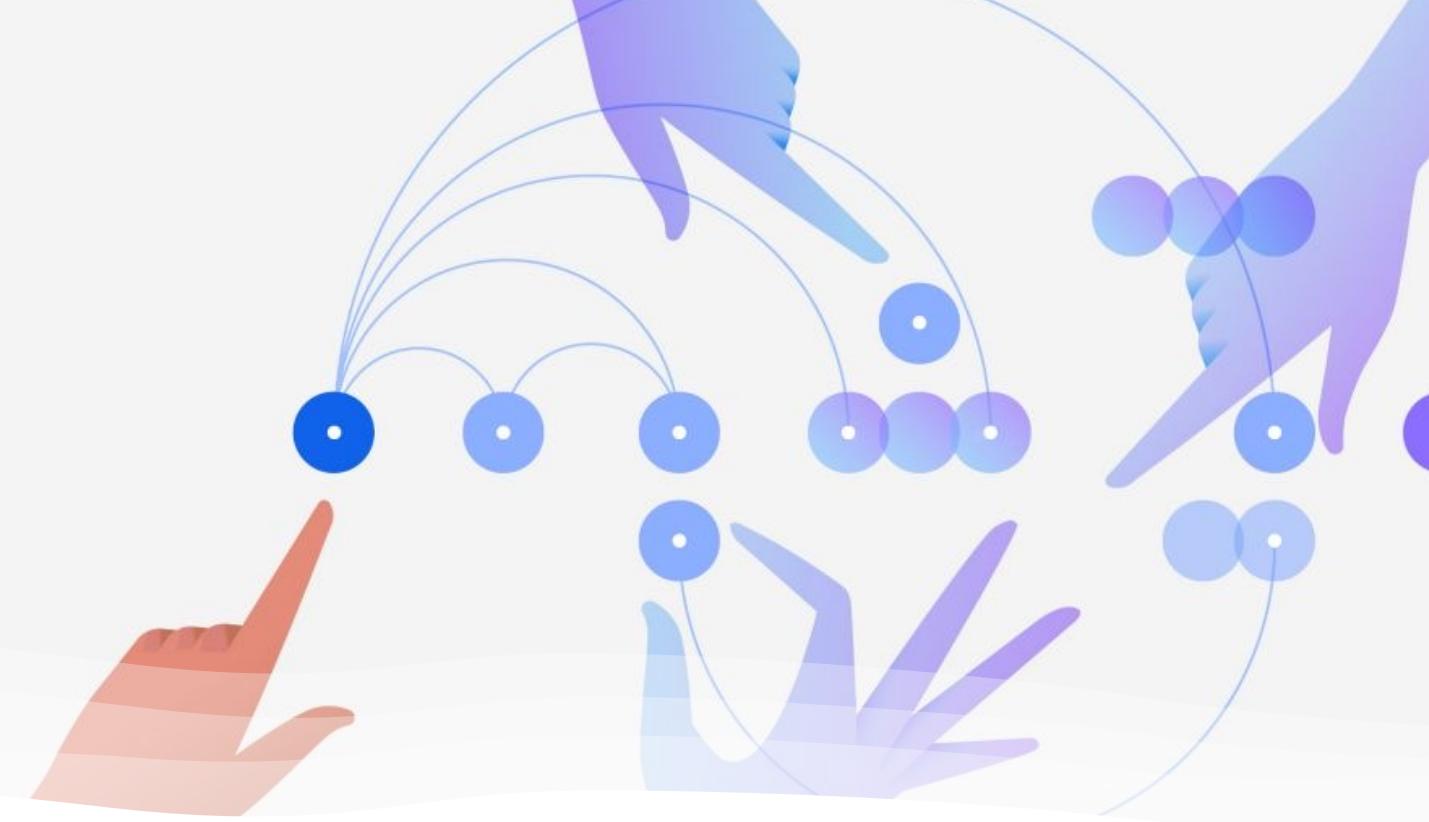
German data

«Mr und Mrs Dursley im Ligusterweg Nummer 4 waren stolz darauf, ganz und gar normal zu sein, sehr stolz sogar.» (Harry Potter)

«Da die Anerkennung der angeborenen Würde und der gleichen und unveräußerlichen Rechte aller Mitglieder der Gemeinschaft der Menschen die Grundlage von Freiheit, Gerechtigkeit und Frieden in der Welt bildet,» (Human Rights)

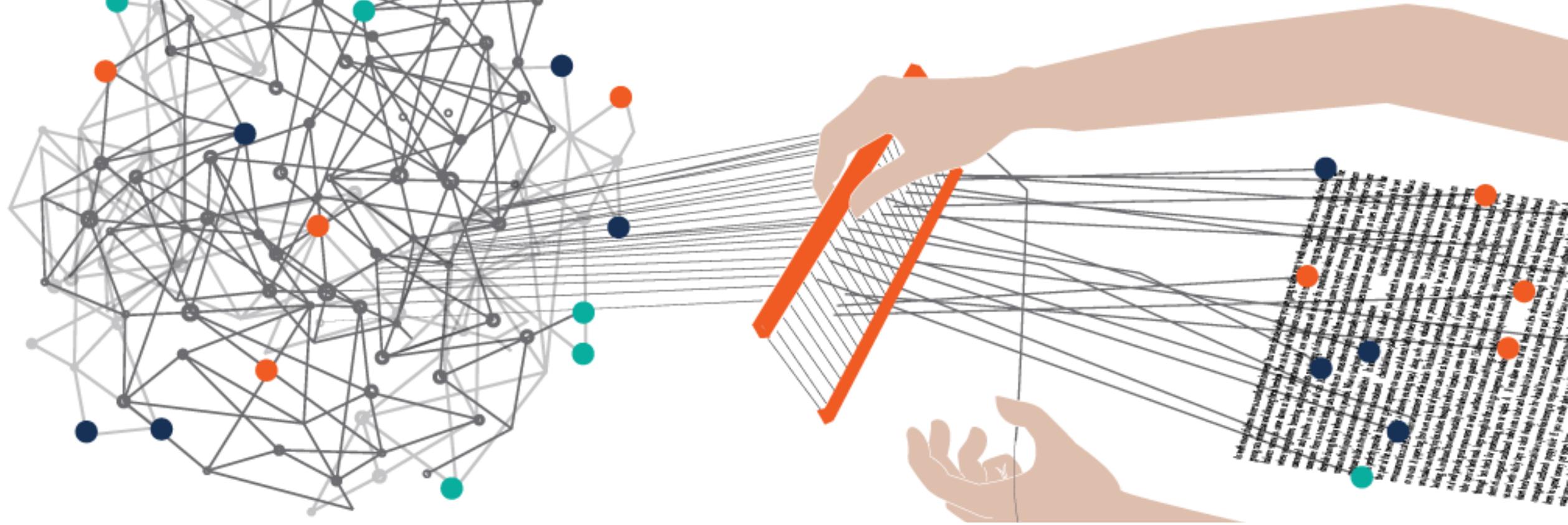
Tokenization and POS Tagging

- Mainly NLTK and SpaCy
- First xml, then csv for easier evaluation
- SpaCy model size varies between languages -> influence on output?



The NLTK POS tagging

```
<w lemma="История" xml:id="Lit_MagicMountain_ru.txt-1_5" pos="NNP">История</w>
<w lemma="Ганса" xml:id="Lit_MagicMountain_ru.txt-1_6" pos="NNP">Ганса</w>
<w lemma="Касторпа" xml:id="Lit_MagicMountain_ru.txt-1_7" pos="NNP">Касторпа</w>
<w lemma="" xml:id="Lit_MagicMountain_ru.txt-1_8" pos=",">,</w>
<w lemma="которую" xml:id="Lit_MagicMountain_ru.txt-1_9" pos="NNP">которую</w>
<w lemma="мы" xml:id="Lit_MagicMountain_ru.txt-1_10" pos="NNP">мы</w>
<w lemma="хотим" xml:id="Lit_MagicMountain_ru.txt-1_11" pos="NNP">хотим</w>
<w lemma="здесь" xml:id="Lit_MagicMountain_ru.txt-1_12" pos="NNP">здесь</w>
```



Semantic role labeling

- Since NLTK performed poorly on previous tasks, we used SpaCy
- SpaCy offers Semantic role labeling
- Results: Relatively accurate

Evaluation

Lemmas			PoS			Dependency		
UK	RU	DE	UK	RU	DE	UK	RU	DE
93.5%	94.72%	92.47%	96.8%	93.65%	93.93%	89.72%	90.4%	91.67%

Sentence alignment

- Alignment with Bertalign
- Only works for German and Russian(?)
- The texts don't have the same content length
- Alignment remains difficult:

Es war noch finster, als er erwachte. Als er nach oben schaute, sah er die Sterne zwischen den Dachbalken durchscheinen.

(„It was still dark when he woke up. As he looked up, he saw the stars shining through the rafters.“

“It was still dark when he woke up. When he looked up, he saw the stars shining through between the roof beams. “)

Он проснулся, когда было еще темно, и сквозь дырявую крышу увидел, как блещут звезды.

(«He woke up when it was still dark, and through the leaky roof he saw the stars shining.»

»He woke up when it was still dark and through the hole in the roof he could see the stars glittering. »)

Next Steps

- Surprisal
- Further evaluation methods?

A group of four stylized human figures standing in a row against a light gray background. Each figure is holding a large, solid-colored question mark above their head. The figures are dressed in simple, modern clothing: two men in orange shirts and blue pants, one man in a black and orange shirt and blue pants, and one woman in an orange shirt and black pants. The question marks are colored black, blue, red, and blue from left to right.

Questions?