PAWUK:
Polish **A**utomatic **W**eb corpus of **UK**rainian

Witold Kieraś **Łukasz Kobyliński** Dorota Komosińska
Bartłomiej Nitoń Michał Rudolf Марія Шведова
Aleksandra Zwierzchowska

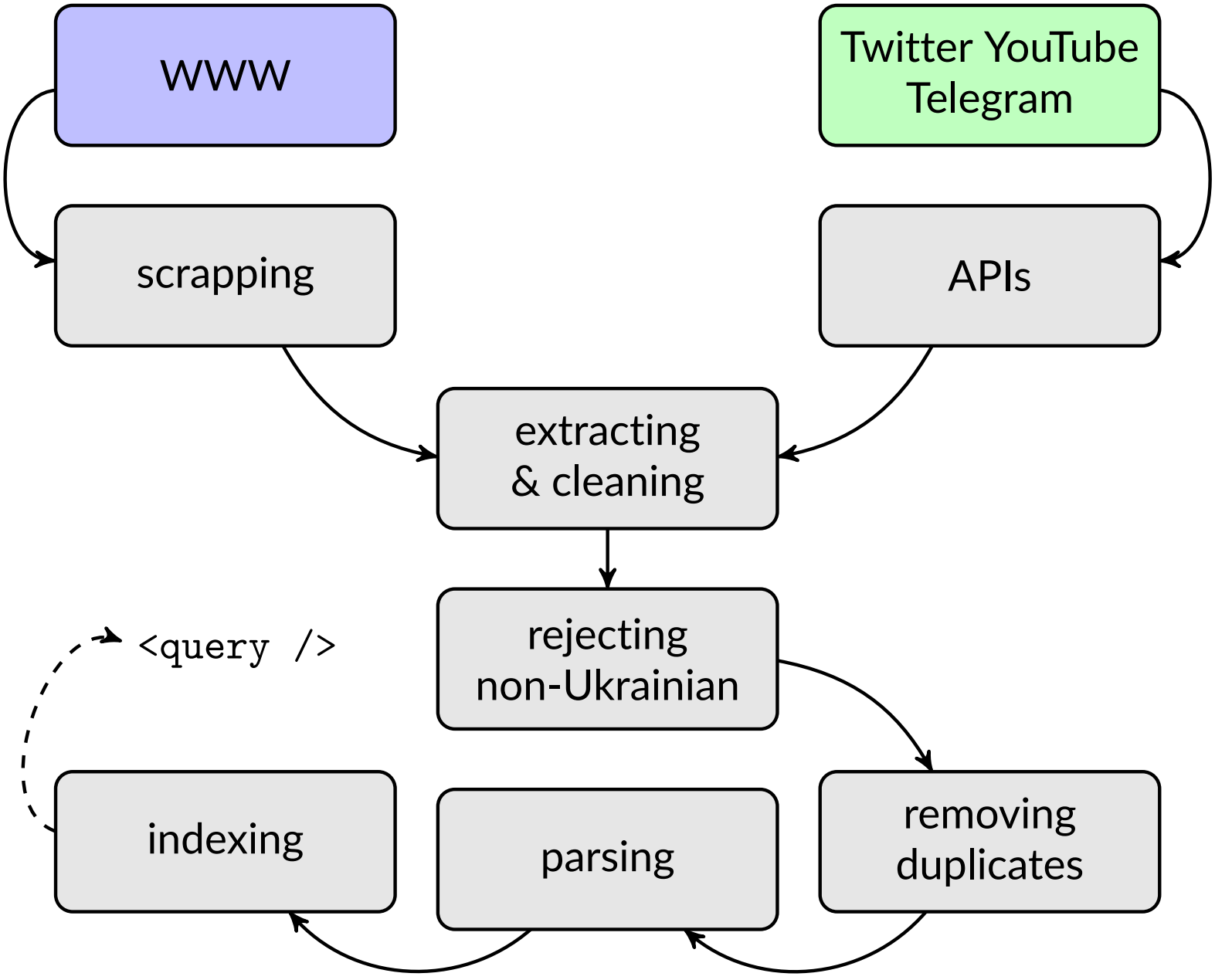
Linguistic Engineering Group
Institute of Computer Science
Polish Academy of Sciences

May 25th 2024, UNLP 2024

Polish Automatic Web corpus of UKrainian

- PAWUK is a Web corpus of Ukrainian created as a collaborative side project with Марія Шведова.
- It monitors selected Ukrainian webpages (200) as well as Ukrainian Twitter (500 accounts), YouTube (300 channels) and Telegram (230 channels).
- Updated on a daily basis (approx. 1.5-2.3 million words everyday; 1.7 on average).
- Currently over 1.3 billion tokens.
- Annotated with a custom Stanza model for:
  - UD morfological tags and features,
  - UD dependency trees,
  - VESUM xpos tags,
  - named entities.
- `https://pawuk.ipipan.waw.pl`

Workflow


WWW
Twitter YouTube Telegram
scrapping
APIs
extracting & cleaning
rejecting non-Ukrainian
<query />
indexing
parsing
removing duplicates

Tagging & parsing

- For annotation processes we use Stanza pipline with custom models for Ukrainian.
- The tagging & parsing models are trained using a sum of two independent resources:
  - standard Ukrainian Universal Dependencies treebank modified to contain ВЕСУМ morphological dictionary tags (known from ГРАК reference corpus) instead of MULTEXT-East tags in `xpos`.
  - Manually annotated with ВЕСУМ tags БрУК[1] (Ukrainian Brown corpus) parsed using the standard Stanza model.
  - The idea behind creating such a heterogeneous training set was to maximize the accuracy of `xpos` tagging without impairing the accuracy of parsing as well as to diversify the training set.
- For named entities the standard Stanza model was used.

[1] `https://github.com/brown-uk`

Evaluation

	UPOS	XPOS	UFeats	AllTags
БрУК parsebank	97.27	92.88	93.15	89.41
UD treebank	98.81	94.38	95.95	91.50
TOTAL	98.29	93.88	95.01	90.80

F1 measure values for testing subsets of the two resources used for training the custom parsing model.

Querying

- orthographic words, ie. `[orth="павуки"]`,
- lemmata: `[lemma="павук"]`,
- UD part of speech: `[upos="NOUN"]`,
- ВЕСУМ tag: `[xpos="noun:anim:p:v_naz"]`,
- UD morphological features: `[ufeat="fem"]`,
- UD dependency relation: `[deprel="nsubj"]`,
- lemma of the syntactic head: `[head.lemma="павук"]`,
- UD part of speech of the syntactic head: `[head.upos="ADJ"]`,
- UD morphological features of the syntactic head: `[head.ufeat="anim"]`,
- named entities: `<ne="PERS"/>`,
- words or morphological interpretations not found in ВЕСУМ: `[oov="true"]`.

Query

[lemma="павук" & deprel="appos"]

QUERY BUILDER | METADATA | STATS

Number of results per page

20

Search

There are 277 results.

No.	Left context	Result	Right context	Channel	Date of publication
1	– малює війну, а не Бетмена чи Людину-	**павука** [павук:noun:anim:m:v_zna]	. Жінка також зауважила, що її здивувало те,	web	2022-03-31
2	Касільяс: Бензема — це Людина-	**Павук** [павук:noun:anim:m:v_naz]	, президент США, янгол-охоронець та Бог!	twitter	2022-04-07
3	місто», «Черкаси», «Людина-	**павук** [павук:noun:anim:m:v_naz]	: Немає шляху додому», «Падіння Місяця»	web	2022-04-21
4	", "Фантастична четвірка", "Людина-	**павук** [павук:noun:anim:m:v_naz]	", "Неймовірний Халк" , "Люди Ікс	web	2022-04-21
5	Spider-Verse 3 тепер називається "Людина-	**павук** [павук:noun:anim:m:v_naz]	: За межами всесвіту" і вийде 29 березня 2024	twitter	2022-04-26
6	переступити заповітну позначку $1 млрд. Тільки «Людина-	**павук** [павук:noun:anim:m:v_naz]	: Додому шляху немає» змогла подолати цей рубіж у	web	2022-05-09
7	що вони самі є владою Згадується фраза з Людини-	**павука**	: “Чим більша сила, тим більша відповідальність“	twitter	2022-05-17