National Technical University
"Kharkiv Polytechnic Institute"

German-Ukrainian Parallel Corpus (ParaRook||DE-UK)

May 25th 2024, UNLP 2024

MARIA SHVEDOVA

ARSENII LUKASHEVSKYI

CURRENT SITUATION

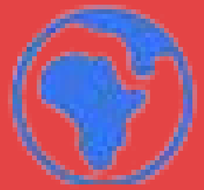
OPUS
OPEN PARALLEL CORPORA

Common Crawl

INTERCORP
projekt paralelních korpusů
Filozofické fakulty Univerzity Karlovy v Praze

лабораторія української

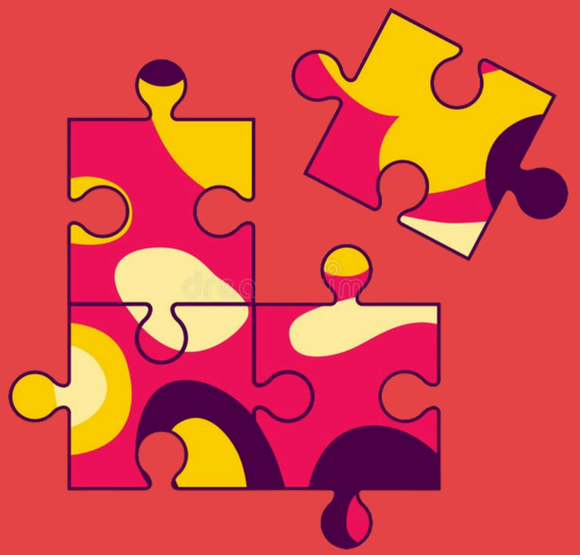
ParaCrawl



No Language Left Behind

CURRENT CHALLENGES

- Limited access to parallel corpora
- Variety of formats
- Fragmentation of projects
- Lack of manually aligned corpora
- Data quality needs improvement

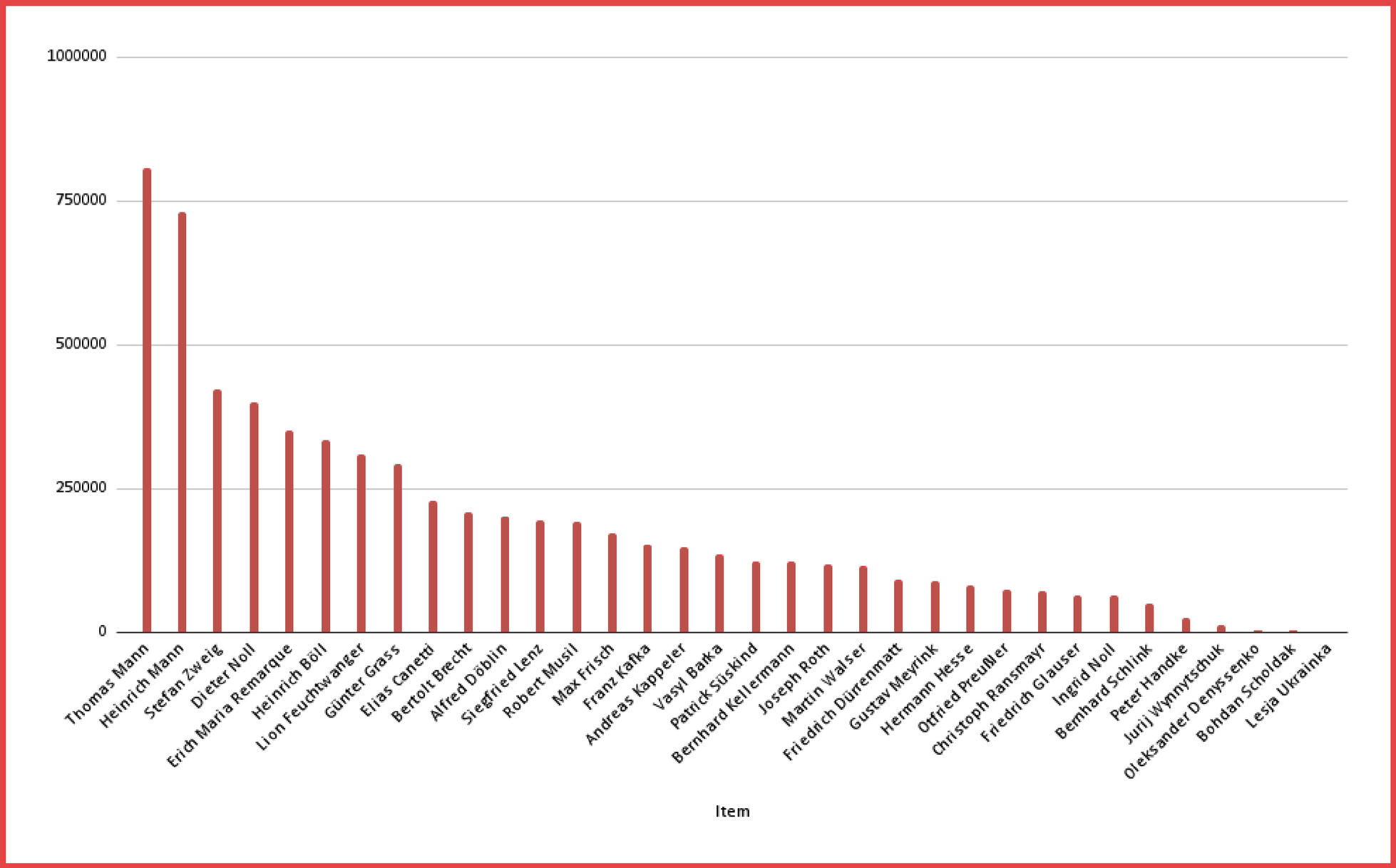


TYPICAL PROBLEMS WITH AUTOMATIC PARALLEL CORPORA

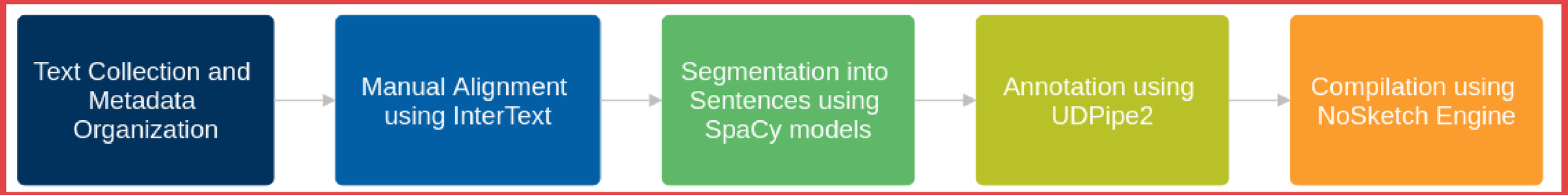
Example of a problem	Example in the original language	Example in the target language
Non-parallel sentences	Der Elektroden Schalter KARI EL22 dient zur Fullstandserfassung und -regelung von elektrisch leitfahigen Fluessigkeiten (The KARI EL22 electrode sensor is used *to detect and control the level* of conductive liquids.) *(Dakwale and Monz, ParaCrawl)*	The KARI EL22 electrode switch is designed for the control of conductive liquids.
Extra sentences	Ви можете вводити ключі пошуку великими або малими літерами, – пошук проводитиметься незалежно від регістру символів. - Щоб провести пошук за декількома ключами пошуку, скористайтесь допоміжним словом OR. Приклад: Вітя OR Юля - Щоб виключити результати з певними ключами з пошуку, поставте перед відповідним ключовим словом знак « мінус », наприклад, - коти - Якщо ви шукаєте ціле словосполучення, візьміть його у лапки. Приклад: « Тут мають бути тигри » - Додайте до ключа пошуку допоміжне словосполучення ext: тип, щоб зазначити суфікс назви файла, наприклад, ext: txt або ext: для файлів без суфіксів. *(OPUS Ukrainian-Estonian, KDE4)*	Mina ja tahaplaanile saatmine otsing on Kellele otsing tingimusel mittekohustuslik VÕI VÕI Kellele otsing asukohas eest kui otsimine tingimusel a fraas lisa Lisa tüüp kuni a fail laiend või tingimusel puudub.
Text defects	E U RO S YS T E M / E S C B C O M M I T T E E S, B U D G E T C O M M I T T E E A N D T H E I R C H A I R P E R S O N S Accounting and Monetary Income Committee (AMICO) Ian Ingram *(OPUS English-German, ECB)*	
Machine translation	Ця людина змусив світ під іншим кутом поглянути на творчість, форму і колір, слова і фрази, моду і її зміст, а також на парфумерне мистецтво. (This person **made* the world look at creativity, form and color, words and phrases, fashion and its meaning, as well as the art of perfumery from a different angle.) (edp.ua; UkTenTen22).	

PARAROOK (DE | | UK)

- ParaRook (DE||UK) is a morphosyntactically annotated parallel corpus of Ukrainian and German texts.
- Morphological and syntactic markup using Universal Dependencies, semi-manual alignment.
- At this point, the main focus is on fiction texts.
- Corpus size: Ukrainian - 6,381,713 tokens, German - 5,847,902 tokens


1000000
750000
500000
250000
0
Thomas Mann
Heinrich Mann
Stefan Zweig
Dieter Noll
Erich Maria Remarque
Heinrich Böll
Lion Feuchtwanger
Günter Grass
Elias Canetti
Bertolt Brecht
Alfred Döblin
Siegfried Lenz
Robert Musil
Max Frisch
Franz Kafka
Andreas Kappeler
Vasyl Barka
Patrick Süskind
Bernhard Kellermann
Joseph Roth
Martin Walser
Friedrich Dürrenmatt
Gustav Meyrink
Hermann Hesse
Otfried Preußler
Christoph Ransmayr
Friedrich Glauser
Ingrid Noll
Bernhard Schlink
Peter Handke
Jurij Wynnytschuk
Oleksander Denyssenko
Bohdan Scholdak
Lesja Ukrainka
Item

WORKFLOW


Text Collection and Metadata Organization
Manual Alignment using InterText
Segmentation into Sentences using SpaCy models
Annotation using UDPipe2
Compilation using NoSketch Engine

ANNOTATION ATTRIBUTES

Attribute	Description
word	Token attribute for a word.
lemma	Token attribute for lemma.
upos	Token attribute for UD part-of-speech tag.
xpos	Language-specific grammatical annotation token attribute.
morphology	Morphological annotation.
head	Syntactically the main word in a sentence.
dependency_tag	Syntactic relationship of a word in a sentence.
extra_dependency	Additional information about the syntactic role of a word in a sentence.
authors_names_{uk|de}, authors_born, authors_sex, authors_regionCode	Authors' name in Ukrainian/German, birth year, gender, and region.
translators_names_{uk|de}, translators_born, translators_sex, translators_regionCode	Translators' name in UK/DE, birth year, gender, and region.
title_{uk|de}	Document title in UK/DE.
original_language	Original language.
date_{uk|de}	Year of creation in UK/DE.
pub_city_{uk|de}, publisher_{uk|de}, pub_year_{uk|de}, publication_{uk|de}	City of publication, publisher, year of publication, and title of publication in UK/DE (magazine number, title of collection).
url_{uk|de}	Reference to the source of the document in UK/DE.

QUERYING EXAMPLE

Роберт Музіль • 1942 • 2010 • Олекса Логвиненко • Людина без власт

Aber während der eine mit diesen Gedanken lächelnd durch den schwebenden Abend ging, hielt der andre die Fäuste geballt, in Schmerz und Zorn; er war der weniger sichtbare, und woran er dachte, war, eine Beschwörungsformel zu finden, einen Griff, den man vielleicht packen könnte, den eigentlichen **Geist des Geistes** , das fehlende, vielleicht nur kleine Stück, das den zerbrochenen Kreis schließt.

Та поки один Ульріх, усміхаючись, простував з такими думками крізь уже завислий над землею вечір, другий з болем і гнівом стискав кулаки; цього було видно не так, як першого, і міркував він про те, щоб знайти якесь заклинання, якийсь важіль, що за нього пощастило б ухопитися, знайти справжній дух духу, якого бракувало, можливо, всього-на-всього невеличкий краєчок, котрий замкне зламаний круг.

Томас Манн • 1924 • 2008 • Роман Осадчук • Зачарована гора

Aber die Lymphe, die ist ja erst der **Saft des Saftes** , die Essenz, wissen Sie, Blutmilch, eine ganz deliziöse … Tropfbarkeit, – nach Fettnahrung sieht sie übrigens wirklich wie Milch aus. "

Але саме лімфа є отим соком усіх соків, квінтесенцією, молоком крови, розумієте, найціннішою рідиною – завдяки посиленому харчуванню жирами вона справді може нагадувати молоко.

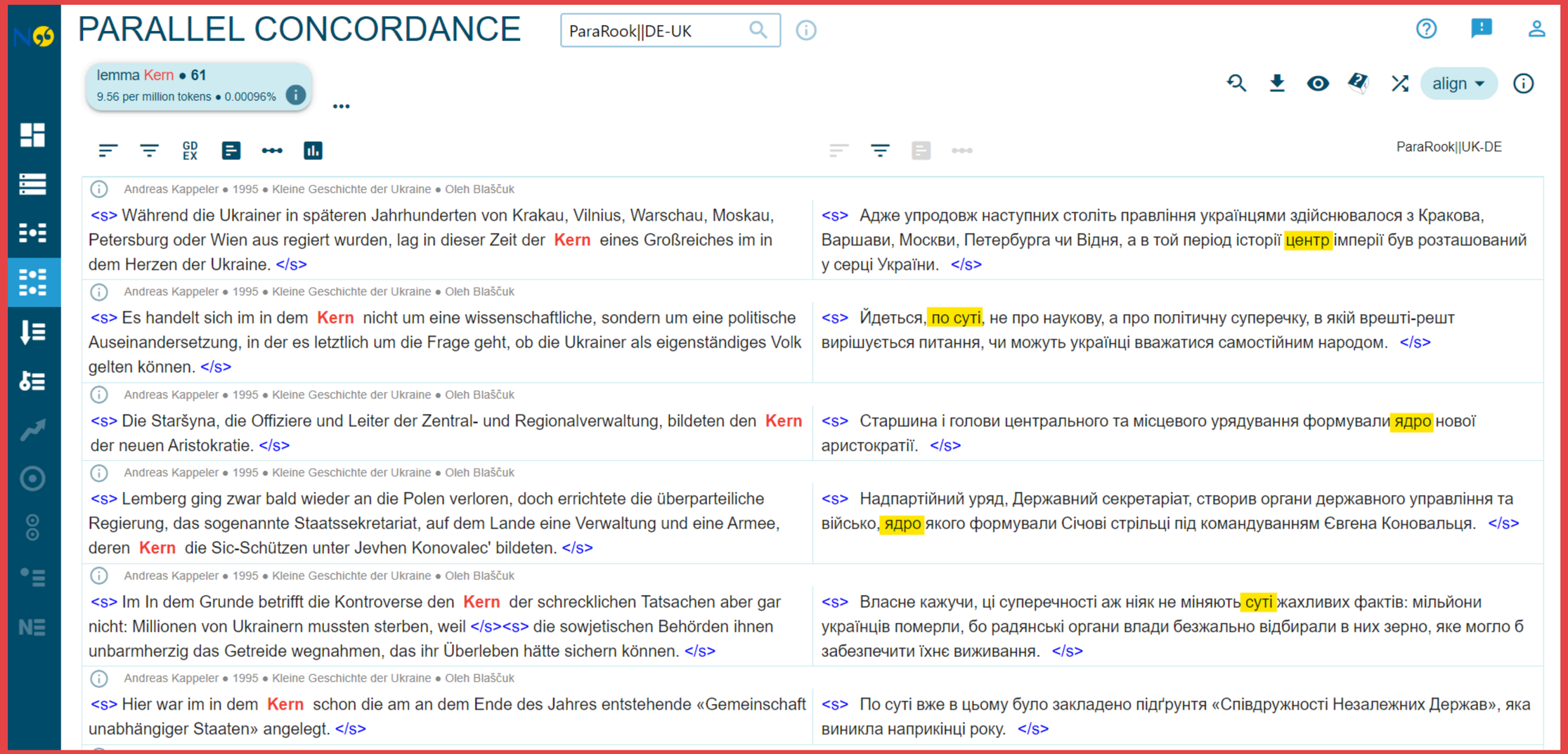
Генріх Манн • 1935 • 1985 • Юрій Лісняк • Літа зрілості короля Генріха

Über den **Punkt der Punkte** dachten sie hinwegzukommen, wär es mit einem Aufgebot von barem Widersinn.

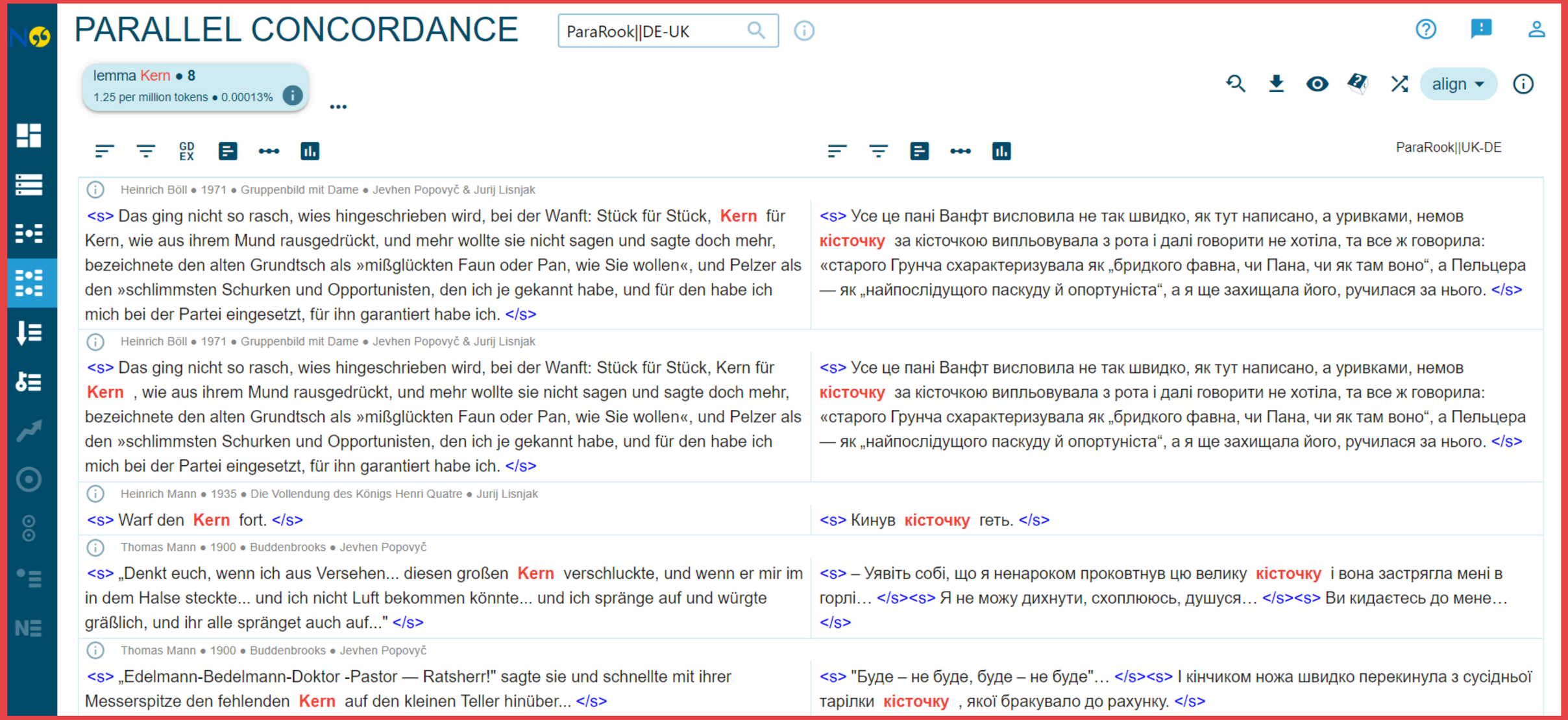
В головному вони сподівались якось викрутитися, хоч би вдавшись до чистісінького безглуздя.

```
1:[upos="N.*"][upos="DET"&word="d.*"]2:[upos="N.*"]&1.lemma=2.lemma
```

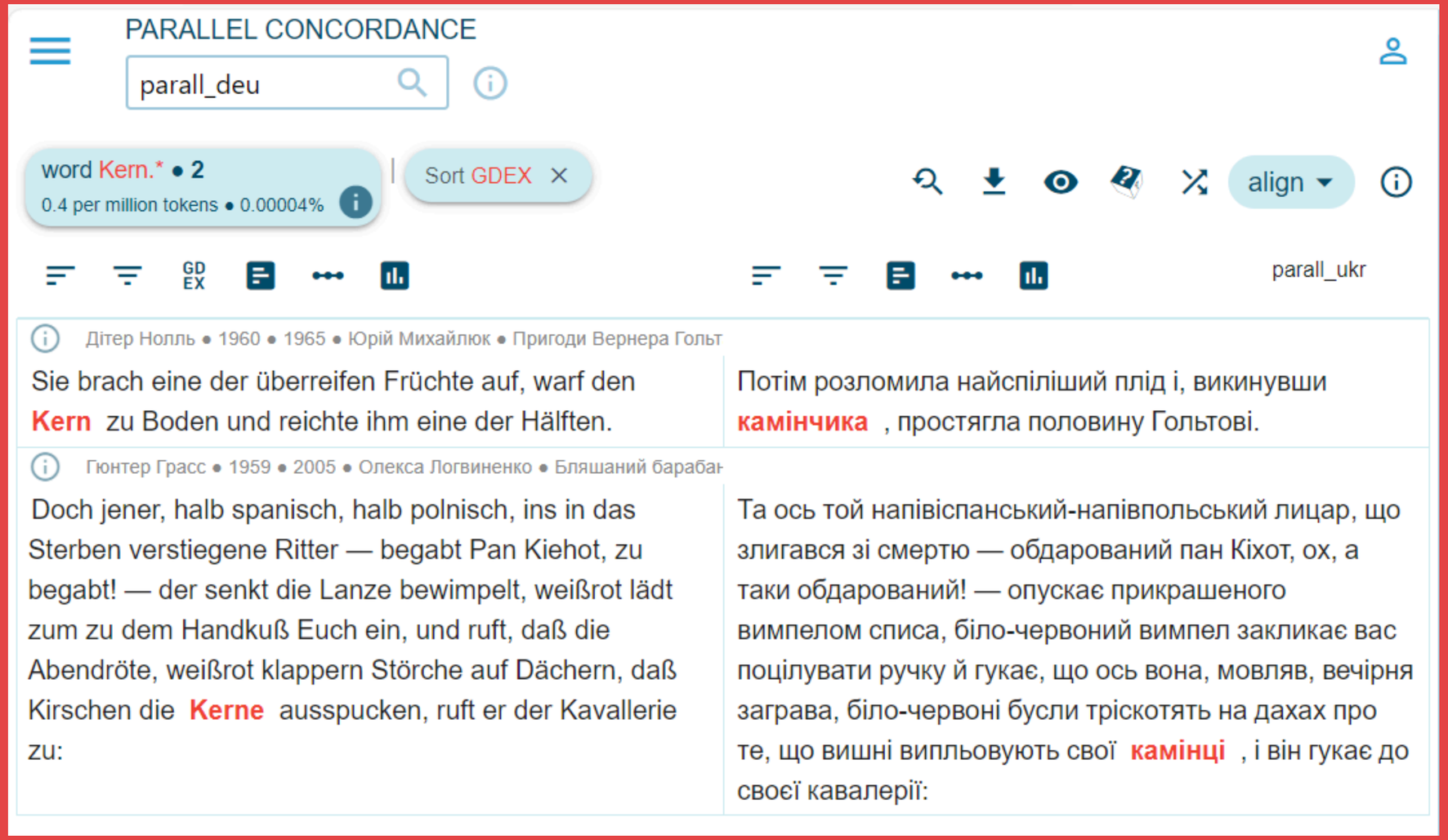
SEMANTIC RESEARCH USING PARAROOK



TRANSLATION VARIATION ANALYSIS



UNUSUAL TRANSLATIONS



CENSORSHIP

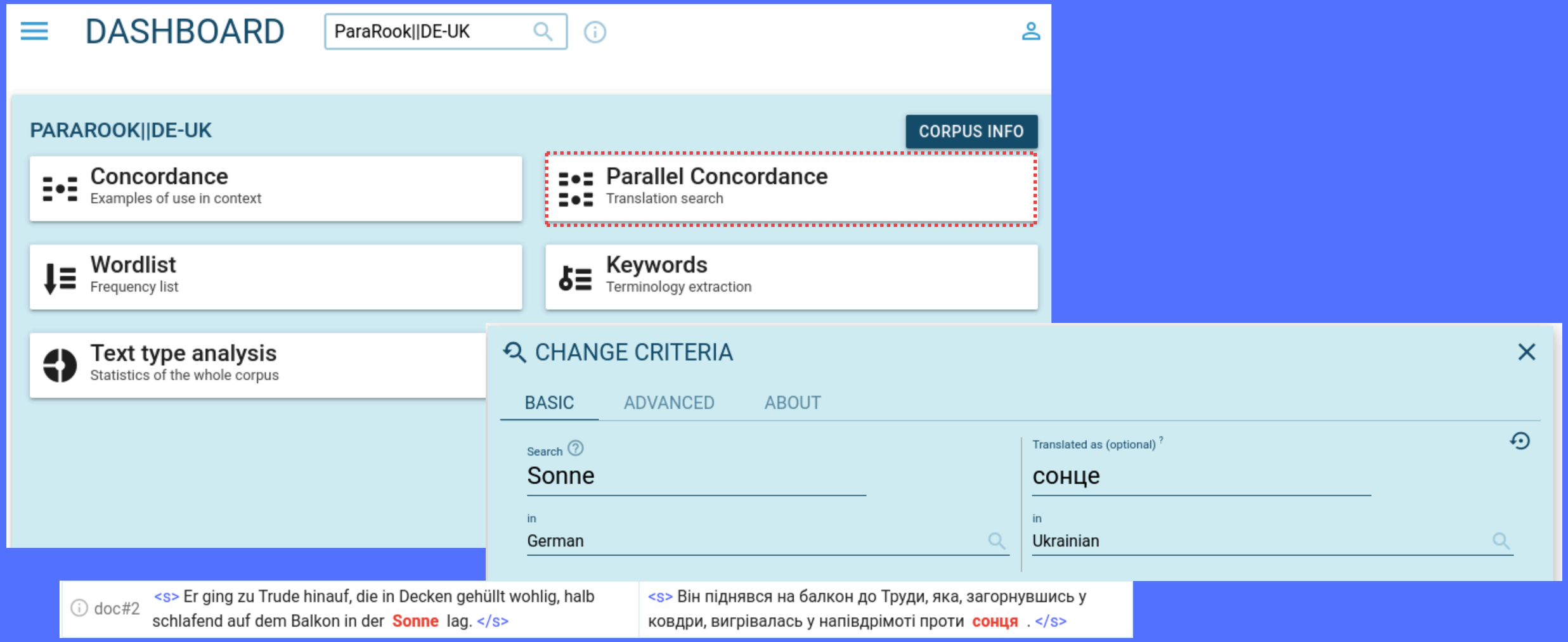
"Цей не може служити Англії, — подумав ґраф.	„Der kann nicht im Dienste Englands stehen’, dachte der Graf.
— У нього, здається, чесне обличчя".	---
І сказав:	Und er fragte:
— Що ж нового, кавалєре?	„Was gibt es Neues, Cavaliere?“

JURIJ KOSSATSCH, “ABEND BEI ROSUMOVSKY”, REGENSBURG, 1946

FUTURE PLANS

- Create a new Ukrainian UD model to improve annotation quality
- Add non-fiction texts to the corpus
- Expand language pairs to include English, French, Spanish, Persian, Japanese, and Chinese
- Develop test corpora for these languages to validate our algorithms across different language pairs

PARAROOK.NYAWKIT.XYZ



THANK YOU