



National Technical University
"Kharkiv Polytechnic Institute"

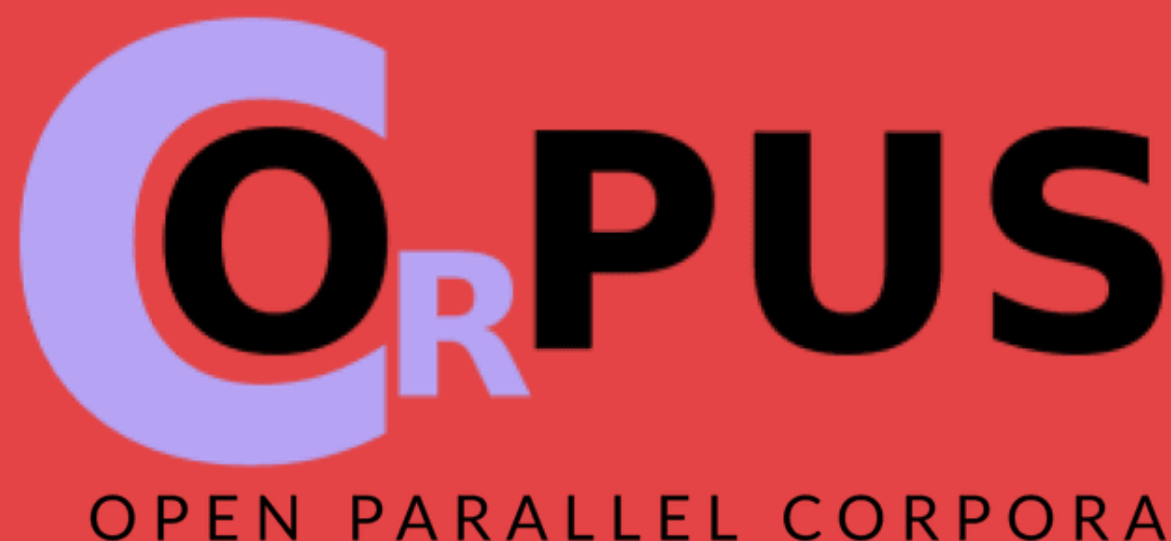
German-Ukrainian Parallel Corpus (ParaRook||DE-UK)

May 25th 2024, UNLP 2024

MARIA SHVEDOVA

ARSENII LUKASHEVSKYI

CURRENT SITUATION

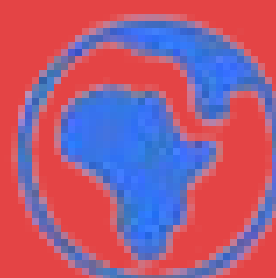


Common
Crawl

INTERCORP
projekt paralelních korpusů
Filozofické fakulty Univerzity Karlovy v Praze

лабораторія української

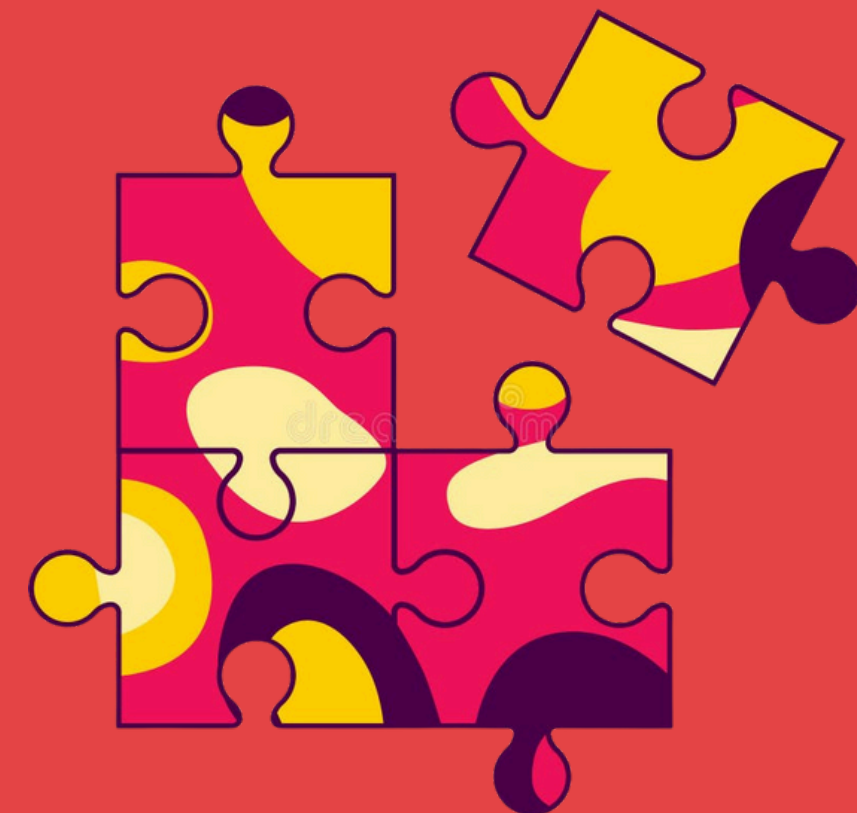
ParaCrawl



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



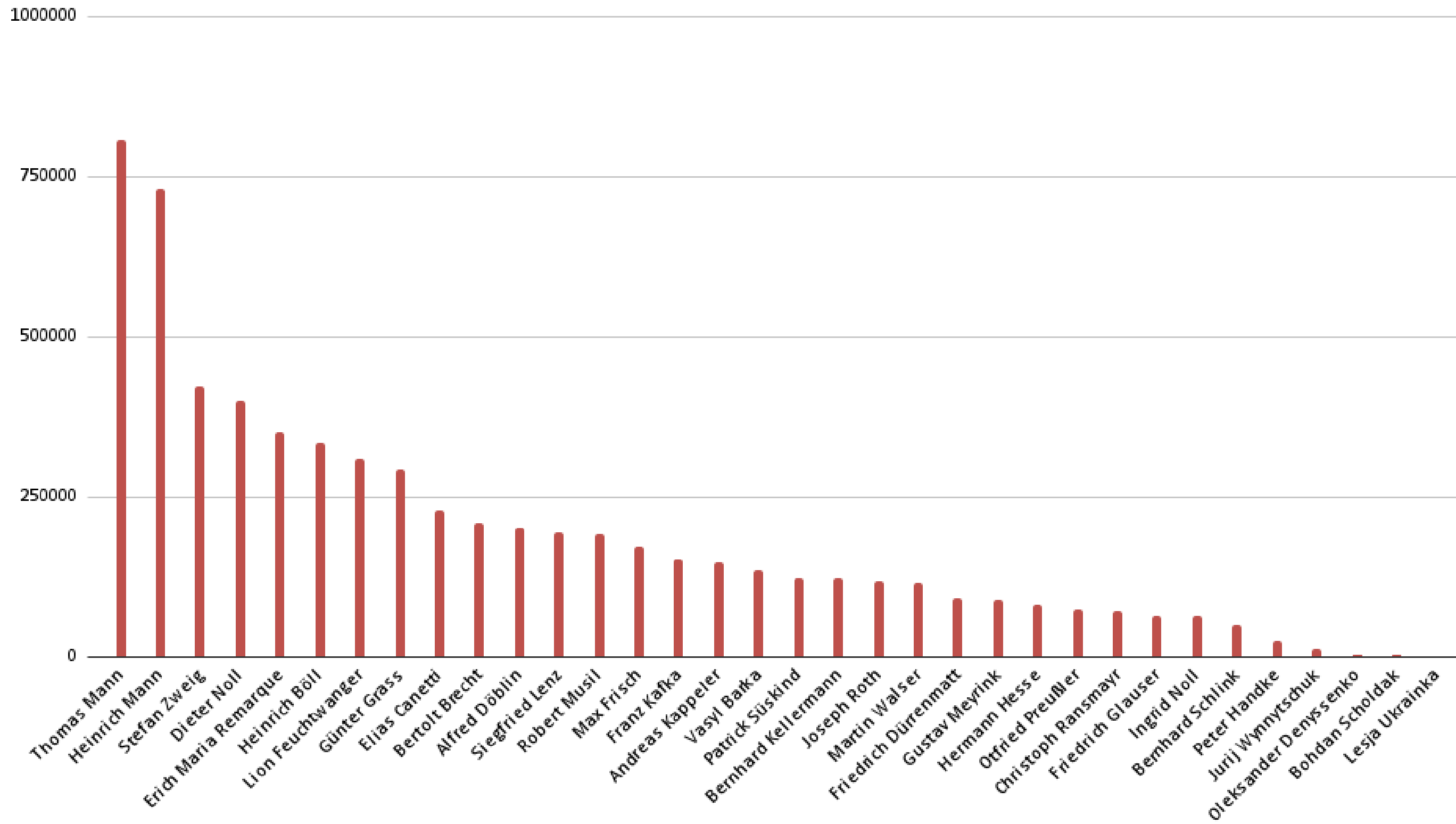
TYPICAL PROBLEMS WITH AUTOMATIC PARALLEL CORPORA

Example of a problem	Example in the original language	Example in the target language
Non-parallel sentences	<p>Der Elektroden Schalter KARI EL22 dient zur Fullstandserfassung und -regelung von elektrisch leitfähigen Flüssigkeiten (The KARI EL22 electrode sensor is used <i>to detect and control the level</i> of conductive liquids.) (Dakwale and Monz, ParaCrawl)</p>	<p>The KARI EL22 electrode switch is designed for the control of conductive liquids.</p>
Extra sentences	<p>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 файлу, наприклад, ext: txt або ext: для файлів без суфіксів. (OPUS Ukrainian-Estonian, KDE4)</p>	<p>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.</p>
Text defects	<p>EURO SYSTEM / ESCB COMMITTEES, BUDGET COMMITTEE AND THEIR CHAIR PERSONS Accounting and Monetary Income Committee (AMICO) Ian Ingram (OPUS English-German, ECB)</p>	
Machine translation	<p>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 <i>*made</i> 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).</p>	

PARAROOK (DE | | UK)

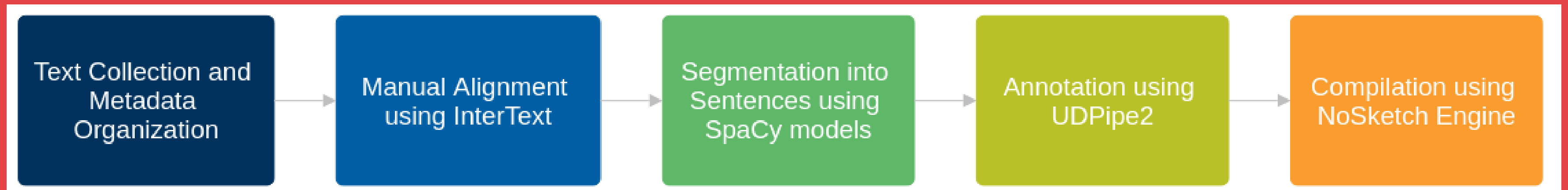
- ParaRook (DE||UK) is a morphosyntactically annotated parallel corpus of Ukrainian and German texts.
- Morphological and syntactic markup using Universal Dependencies, semi-manual alignment.
- At this point, the main focus is on fiction texts.
- Corpus size: Ukrainian - 6,381,713 tokens, German - 5,847,902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





Item

WORKFLOW



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

<p>i Роберт Музіль • 1942 • 2010 • Олекса Логвиненко • Людина без власт</p> <p>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.</p>	<p>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 зламаный круг.</p>
<p>i Томас Манн • 1924 • 2008 • Роман Осадчук • Зачарована гора</p> <p>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. "</p>	<p>Але саме лімфа є отим соком усіх соків, квінтесенцією, молоком крові, розумієте, найціннішою рідиною – завдяки посиленому харчуванню жирами вона справді може нагадувати молоко.</p>
<p>i Генріх Манн • 1935 • 1985 • Юрій Лісняк • Літа зрілості короля Генріха</p> <p>Über den Punkt der Punkte dachten sie hinwegzukommen, wär es mit einem Aufgebot von barem Widersinn.</p>	<p>В головному вони сподівались якось викрутитися, хоч би вдавшись до чистісінького безглуздя.</p>

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

SEMANTIC RESEARCH USING PARAROOK

PARALLEL CONCORDANCE ParaRook||DE-UK

lemma **Kern** • 61
9.56 per million tokens • 0.00096%

align

ParaRook||UK-DE

<p>Andreas Kappeler • 1995 • Kleine Geschichte der Ukraine • Oleh Blaščuk</p> <p><s> Während die Ukrainer in späteren Jahrhunderten von Krakau, Vilnius, Warschau, Moskau, Petersburg oder Wien aus regiert wurden, lag in dieser Zeit der Kern eines Großreiches im in dem Herzen der Ukraine. </s></p>	<p><s> Адже упродовж наступних століть правління українцями здійснювалося з Кракова, Варшави, Москви, Петербурга чи Відня, а в той період історії центр імперії був розташований у серці України. </s></p>
<p>Andreas Kappeler • 1995 • Kleine Geschichte der Ukraine • Oleh Blaščuk</p> <p><s> Es handelt sich im in dem Kern nicht um eine wissenschaftliche, sondern um eine politische Auseinandersetzung, in der es letztlich um die Frage geht, ob die Ukrainer als eigenständiges Volk gelten können. </s></p>	<p><s> Йдеться, по суті, не про наукову, а про політичну суперечку, в якій врешті-решт вирішується питання, чи можуть українці вважатися самостійним народом. </s></p>
<p>Andreas Kappeler • 1995 • Kleine Geschichte der Ukraine • Oleh Blaščuk</p> <p><s> Die Staršyna, die Offiziere und Leiter der Zentral- und Regionalverwaltung, bildeten den Kern der neuen Aristokratie. </s></p>	<p><s> Старшина і голови центрального та місцевого урядування формували ядро нової аристократії. </s></p>
<p>Andreas Kappeler • 1995 • Kleine Geschichte der Ukraine • Oleh Blaščuk</p> <p><s> Lemberg ging zwar bald wieder an die Polen verloren, doch errichtete die überparteiliche Regierung, das sogenannte Staatssekretariat, auf dem Lande eine Verwaltung und eine Armee, deren Kern die Sic-Schützen unter Jevhen Konovalc' bildeten. </s></p>	<p><s> Надпартійний уряд, Державний секретаріат, створив органи державного управління та військо, ядро якого формували Січові стрільці під командуванням Євгена Коновальця. </s></p>
<p>Andreas Kappeler • 1995 • Kleine Geschichte der Ukraine • Oleh Blaščuk</p> <p><s> Im In dem Grunde betrifft die Kontroverse den Kern der schrecklichen Tatsachen aber gar nicht: Millionen von Ukrainern mussten sterben, weil </s><s> die sowjetischen Behörden ihnen unbarmerzig das Getreide wegnahmen, das ihr Überleben hätte sichern können. </s></p>	<p><s> Власне кажучи, ці суперечності аж ніяк не міняють суті жахливих фактів: мільйони українців померли, бо радянські органи влади безжально відбирали в них зерно, яке могло б забезпечити їхнє виживання. </s></p>
<p>Andreas Kappeler • 1995 • Kleine Geschichte der Ukraine • Oleh Blaščuk</p> <p><s> Hier war im in dem Kern schon die am an dem Ende des Jahres entstehende «Gemeinschaft unabhängiger Staaten» angelegt. </s></p>	<p><s> По суті вже в цьому було закладено підґрунтя «Співдружності Незалежних Держав», яка виникла наприкінці року. </s></p>

TRANSLATION VARIATION ANALYSIS

PARALLEL CONCORDANCE ParaRook||DE-UK

lemma **Kern** • 8
1.25 per million tokens • 0.00013%

align

ParaRook||UK-DE

<p>Heinrich Böll • 1971 • Gruppenbild mit Dame • Jevhen Popovyč & Jurij Lisnjak</p> <p><s> Das ging nicht so rasch, wies hingeschrieben wird, bei der Wanft: Stück für Stück, Kern für Kern, wie aus ihrem Mund rausgedrückt, und mehr wollte sie nicht sagen und sagte doch mehr, bezeichnete den alten Grundtsch als »mißglückten Faun oder Pan, wie Sie wollen«, und Pelzer als den »schlimmsten Schurken und Opportunisten, den ich je gekannt habe, und für den habe ich mich bei der Partei eingesetzt, für ihn garantiert habe ich. </s></p>	<p><s> Усе це пані Ванфт висловила не так швидко, як тут написано, а уривками, немов кісточку за кісточкою випльовувала з рота і далі говорити не хотіла, та все ж говорила: «старого Грунча схарактеризувала як „бридкого фавна, чи Пана, чи як там воно“, а Пельцера — як „найпослідушого паскуду й опортуніста“, а я ще захищала його, ручилася за нього. </s></p>
<p>Heinrich Böll • 1971 • Gruppenbild mit Dame • Jevhen Popovyč & Jurij Lisnjak</p> <p><s> Das ging nicht so rasch, wies hingeschrieben wird, bei der Wanft: Stück für Stück, Kern für Kern , wie aus ihrem Mund rausgedrückt, und mehr wollte sie nicht sagen und sagte doch mehr, bezeichnete den alten Grundtsch als »mißglückten Faun oder Pan, wie Sie wollen«, und Pelzer als den »schlimmsten Schurken und Opportunisten, den ich je gekannt habe, und für den habe ich mich bei der Partei eingesetzt, für ihn garantiert habe ich. </s></p>	<p><s> Усе це пані Ванфт висловила не так швидко, як тут написано, а уривками, немов кісточку за кісточкою випльовувала з рота і далі говорити не хотіла, та все ж говорила: «старого Грунча схарактеризувала як „бридкого фавна, чи Пана, чи як там воно“, а Пельцера — як „найпослідушого паскуду й опортуніста“, а я ще захищала його, ручилася за нього. </s></p>
<p>Heinrich Mann • 1935 • Die Vollendung des Königs Henri Quatre • Jurij Lisnjak</p> <p><s> Warf den Kern fort. </s></p>	<p><s> Кинув кісточку геть. </s></p>
<p>Thomas Mann • 1900 • Buddenbrooks • Jevhen Popovyč</p> <p><s> „Denkt euch, wenn ich aus Versehen... diesen großen Kern verschluckte, und wenn er mir im in dem Halse steckte... und ich nicht Luft bekommen könnte... und ich spränge auf und würgte gräßlich, und ihr alle spränget auch auf...“ </s></p>	<p><s> – Уявіть собі, що я ненароком проковтнув цю велику кісточку і вона застрягла мені в горлі... </s><s> Я не можу дихнути, схоплююсь, душуся... </s><s> Ви кидаєтесь до мене... </s></p>
<p>Thomas Mann • 1900 • Buddenbrooks • Jevhen Popovyč</p> <p><s> „Edelmann-Bedelmann-Doktor -Pastor — Ratsherr!“ sagte sie und schnellte mit ihrer Messerspitze den fehlenden Kern auf den kleinen Teller hinüber... </s></p>	<p><s> "Буде – не буде, буде – не буде"... </s><s> І кінчиком ножа швидко перекинула з сусідньої тарілки кісточку , якої бракувало до рахунку. </s></p>

UNUSUAL TRANSLATIONS

PARALLEL CONCORDANCE

parall_deu

word **Kern.*** • 2
0.4 per million tokens • 0.00004%

Sort **GDEX** ×

align ▾

GD EX

parall_ukr

Дітер Нолль • 1960 • 1965 • Юрій Михайлюк • Пригоди Вернера Гольт

Sie brach eine der überreifen Früchte auf, warf den **Kern** zu Boden und reichte ihm eine der Hälften.

Потім розломилла найспіліший плід і, викинувши **камінчика**, простягла половину Гольтові.

Гюнтер Грасс • 1959 • 2005 • Олекса Логвиненко • Бляшаний барабан

Doch jener, halb spanisch, halb polnisch, ins in das Sterben verstiegene Ritter — begabt Pan Kiehot, zu begabt! — der senkt die Lanze bewimpelt, weißrot lädt zum zu dem Handkuß Euch ein, und ruft, daß die Abendröte, weißrot klappern Störche auf Dächern, daß Kirschen die **Kerne** ausspucken, ruft er der Kavallerie zu:

Та ось той напівіспанський-напівпольський лицар, що злигався зі смертю — обдарований пан Кіхот, ох, а таки обдарований! — опускає прикрашеного вимпелом списа, біло-червоний вимпел закликає вас поцілувати ручку й гукає, що ось вона, мовляв, вечірня заграва, біло-червоні бусли тріскотять на дахах про те, що вишні випльовують свої **камінці**, і він гукає до своєї кавалерії:

CENSORSHIP

▶ "Цей не може служити Англії, — подумав граф.	▶ „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

☰ DASHBOARD ⓘ

PARAROOK||DE-UK CORPUS INFO

- Concordance**
Examples of use in context
- Parallel Concordance**
Translation search
- Wordlist**
Frequency list
- Keywords**
Terminology extraction
- Text type analysis**
Statistics of the whole corpus

CHANGE CRITERIA ⓘ

BASIC | ADVANCED | ABOUT

Search ⓘ
Sonne

in
German

Translated as (optional) ? ⓘ
сонце

in
Ukrainian

ⓘ doc#2

<s> Er ging zu Trude hinauf, die in Decken gehüllt wohlig, halb schlafend auf dem Balkon in der **Sonne** lag. </s>

<s> Він піднявся на балкон до Труді, яка, загорнувшись у ковдри, вигрівалась у напівдрімоті проти **сонця** . </s>

THANK YOU