



# BrUK Group for Ukrainian NLP

Tools and Data



# Team



**Andriy  
Rysin**

Principal  
Software  
Developer  
**SAS Institute  
Inc.**



**Vasyl  
Starko**

PhD in  
Linguistics,  
Independent  
researcher



**Mariana  
Romanyshyn**

Technical Lead, Sr.  
Computational  
Linguist  
**Grammarly, Inc.**



**Nataliia  
Cheilytko**

NLP Practice  
Leader, PhD  
**Envion  
Software**



**Nastasiya  
Osidach**

Engineering  
Manager,  
**Grammarly,  
Inc.**



**Olga  
Havura**

Linguist, PhD

**... and many others**

# LanguageTool for Ukrainian [languagetool.org/uk](https://languagetool.org/uk)

УВАГА! Внизу наведено приклад тексту з помилками, які допоможе виправити LanguageTool. Будь-ласка, вставте тут ваш текст, або перевірте цей текст на предмет помилок. Знайти всі помилки для LanguageTool є не по силах з багатьох причин але дещо він вам все таки підкаже. Порівняно з засобами перевірки орфографії LanguageTool також змайде граматичні та стилюві проблеми. LanguageTool – ваш самий кращий помічник.

українська ▾

Перевірити

LanguageTool – вільний програмний засіб для перевірки граматики та стилю для української мови, також підтримує → 25 інших мов.

Спільнота LanguageTool  
Правила помилок для LanguageTool

Breton Catalan Dutch English Esperanto French German Polish Portuguese Russian Spanish Ukrainian ▾

## Аналіз тексту

Показати результати внутрішнього аналізу LanguageTool, що дають змогу зрозуміти, на основі чого спрацьовують його правила:

Сьогодні ми провели ще одну зустріч.



Проаналізувати текст Підказка: також можна відправити цю форму через Ctrl+Return

## Analysis Result

LanguageTool version: 3.6-SNAPSHOT (2016-10-07 22:01)

Language: Ukrainian

### What do the tags mean?

Disambiguator log: (no disambiguations)

Token	Lemma	Part-of-speech	Chunk
	-	SENT_START	
Сьогодні	сьогодні	adv	
ми	ми	noun:anim:р:v_naz:&pron:pers:1	
провели	проводити	verb:perf:past:p	
щє	щє	adv part	
одну	один	adj:fv_zna:&pron:dem:ind numr:fv_zna	
зустріч	зустріч	noun:inanim:f:v_naz noun:inanim:f:v_zna	
.	-	SENT_END PARA_END	

## Also via REST API

<http://nlp.net.ua:8787/languagetool/swagger-ui.html>

[http://nlp.net.ua:8787/dict\\_uk/swagger-ui.html](http://nlp.net.ua:8787/dict_uk/swagger-ui.html)

# Latest version of LanguageTool 3.5 for Ukrainian

- dictionary with over 208,000 lemmas  
BECYM: [brown-uk/dict\\_uk](#)
- more than 370 grammar and stylistic rules (work on all wordforms)
- over 2,400 word choice suggestions

# Challenge: Machine-translated text

The screenshot shows the homepage of the 112 Ukraine news website. At the top, there is a navigation bar with links for 'ХРОНІКИ 112', 'АРХІВ', 'ТЕЛЕКАНАЛ', a search bar with placeholder text 'Ві пошук...', and a red button for 'ПРЯМІЙ ЕФІР "112"'. Below the navigation bar, there is a menu with categories: 'экономіка', 'Суспільство', 'НП', 'Афіша Києва', 'Київ', 'Спорт', and 'Рада онлайн'. On the right side of the header, there are buttons for 'дивитися' (watch) and 'СЛУХАТИ' (listen). The main content area features a large headline in Ukrainian: 'На вулицях Анкари почалися перейми військових з поліцією, президентський палац без охорони'. Below the headline, there is a timestamp '00:05, 16 липня 2016' and sharing options for Facebook ('Поделиться 9') and Twitter ('Твітнути').

**The military and the police went into labor pains in the streets of Ankara ???**

# Challenge: Disambiguation

Великі **дані** vs. **дані** проекти

За **даними** уряду vs. за **даними** фактами

База **даних** vs. у **даних** умовах

# Challenges accepted

- Data acquisition and preparation
- Training Word2Vec for Ukrainian
- Building 2 Random Forest classifiers to:
  - solve the **дані** case
  - detect unedited machine-translated texts

# Lessons Learned

Given the results obtained on 2.5-mln-token set lemmatized and 300-sentence labeled set per case:

- **Need for data** - a good Ukrainian corpus
- Improved data preprocessing is a must
- Word2Vec works, but there are other options to be explored
- Tokens work, but character-based vectors are well worth considering

# Ukrainian Brown Corpus

Create a corpus of modern Ukrainian based on the principles of the Brown Corpus:

- balanced
- annotated
- freely available

# Main principles

1. Original (not translated) and edited texts.
2. Created and published within a fairly brief period of time.
3. Texts selected following roughly the classification matrix of the original Brown Corpus.
4. Texts represent a variety of sources, authors, topics, and genres.

# BrUK features

- 1 mln tokens
- 9 categories (from newspaper texts to fiction)
- 500+ fragments
- annotated (PoS tags)
- disambiguated (homonyms removed)
- dictionary + tagger

# No more surzhik!

"наростаюча паніка"

"несуча стіна"

"відділення" в значенні "відділ"/"відділок"

"вірний" в значенні "правильний"

"постановка" в значенні "постанова" (театральної вистави)

"принадлежність" замість "належність"



How about a mobile app  
for the Ukrainian  
community to  
**LIKE AND SHARE**  
TEXTS  
**IN GOOD UKRAINIAN?**

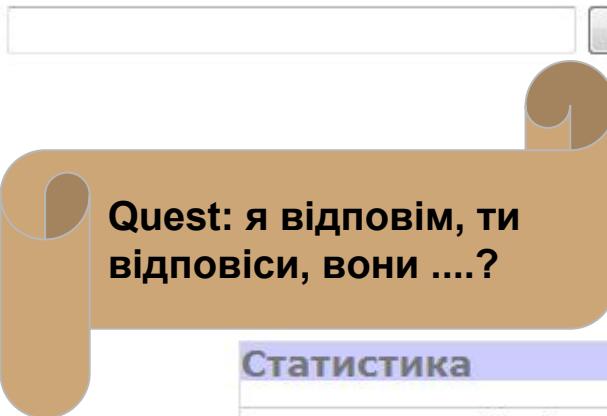
# Next steps

- finish collecting texts for BrUK
- annotate the corpus (automatic PoS tagging + manual verification)
- train various ML models (for PoS tagging, semantic similarity detection, error correction, etc.)
- collect bigger general corpora, specialized corpora, etc.

# Keep in mind: [r2u.org.ua](http://r2u.org.ua)

## Російсько-українські словникі

Про пошук | Словники | Перевірити текст | Статті  
Посилання | Правопис | Переклади | Форум | e2u.org.ua



Знайти

Серед всіх слів

Форматування:

- Сховати наголоси  
 Освітлювати знайдене

У всіх словниках

У всіх словниках

1. Кримський. Академічний рос-укр
2. Ізюмов. Російсько-український
3. Підмогильний. Фразеологічний
4. Шелудько. Тех. термінологі
5. Дорошенко. Ділової мови
6. Вирган, Пилинська. Сталих виразів
7. Ніковський. Українсько-російський
8. Уманець, Спілка. Словар
9. Народний рос-укр словник
10. Голоскевич. Правописний словник
11. Грінченко. Словар' української мови
12. Якубські. Військової термінологі
13. Кобів. Словник судинних рослин
14. Народний тлумачно-стилістичний
15. Младзинський. Словник приказок

### Статистика

Всього статей в базі: 345 524  
унікальних статей: 197 232  
статей в народному р-у: 4 688  
статей в тлумачному: 217

### Новини

# Keep in mind: e2u.org.ua

Словники | Переклади | Форум | r2u.org.ua

## Англійсько-українські словники

Знайти

У всіх словниках

- У всіх словниках
- Загальний англо-український словник
- Словник Мейнаровича, Кратка
- Англо-український словник з ІТ
- Англо-український словник наукової мови**
- Українсько-англійський словник наукової мови
- Словник термінології Європейського Союзу
- Довідник ідіом і виразів
- Українсько-англійський словник
- Фразлекс
- Укр-англ. словник лінгвістичної термінології
- Укр-англ. словник з прав людини

Параметри пошуку:

- Лише серед головних слів
- Освітлювати знайдене

New Ukrainian words, e.g.  
debug -  
зневаджувати

(вада ->  
зневаждувати  
як  
шкода ->  
знешкоджувати)

### Статистика

Всього статей в базі: 172 475  
унікальних статей: 160 281  
статей в народному: 4 112

### Оголошення

16.05.2016

В базу сайту додано «Українсько-англійський словник з прав людини» Леся Герасимчук 2015 року видання

# Potential users and contributors

- software engineers
- NLP practitioners
- NLP researchers
- ML engineers
- data scientists
- AI specialists
- linguists
- etc.



# Eager to join in ?

**Contact us via email:** [bruk.group@gmail.com](mailto:bruk.group@gmail.com)

**GitHub:** [Brown-uk](#)  
[nlp\\_uk](#)  
[dict\\_uk](#)  
[corpus](#)

**Find more about us:** <http://r2u.org.ua/corpus>

**Facebook:** [www.facebook.com/r2u.org.ua/](http://www.facebook.com/r2u.org.ua/)

Thank you, you are the best!

Any questions?