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

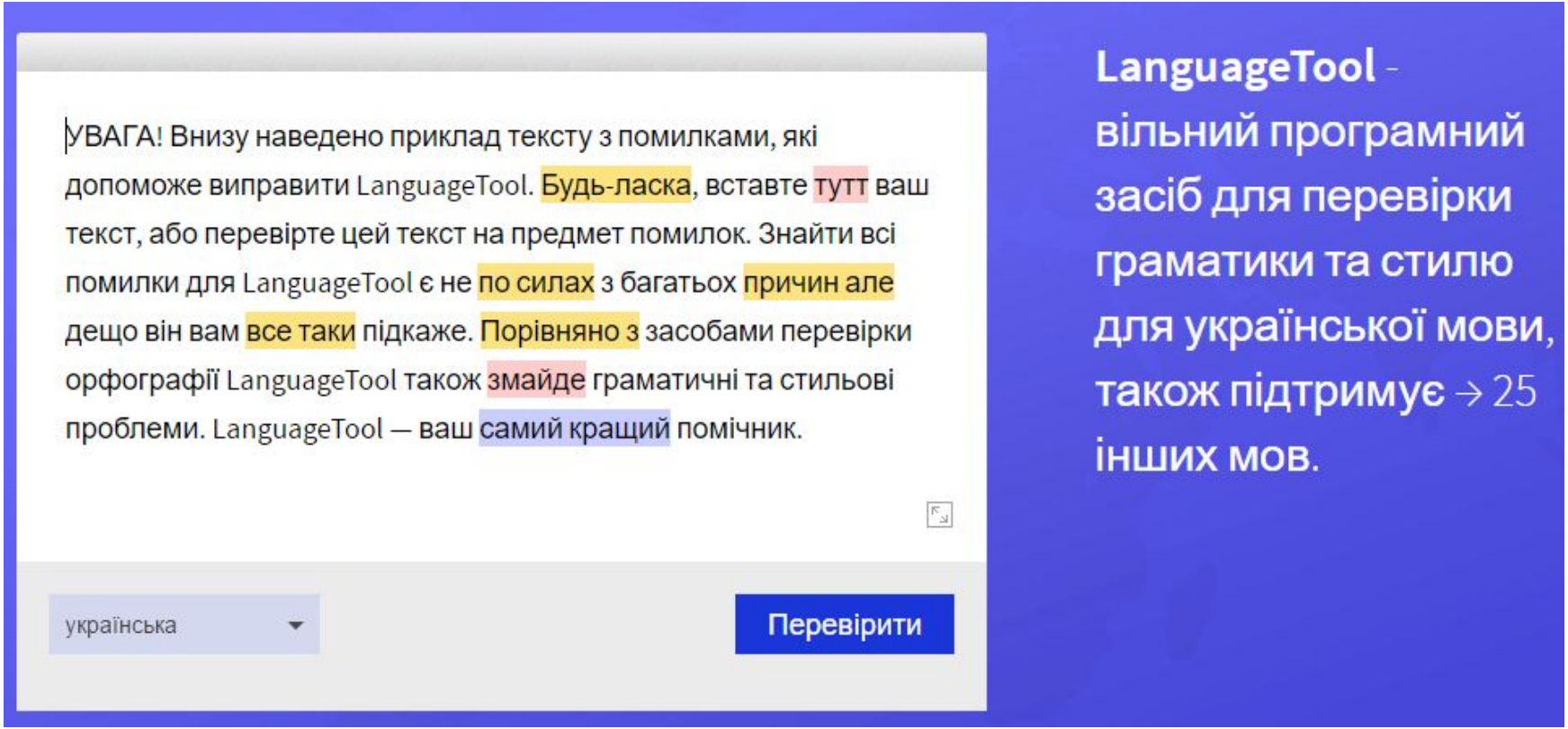
Engineering Manager, **Grammarly, Inc.**

Olga Havura

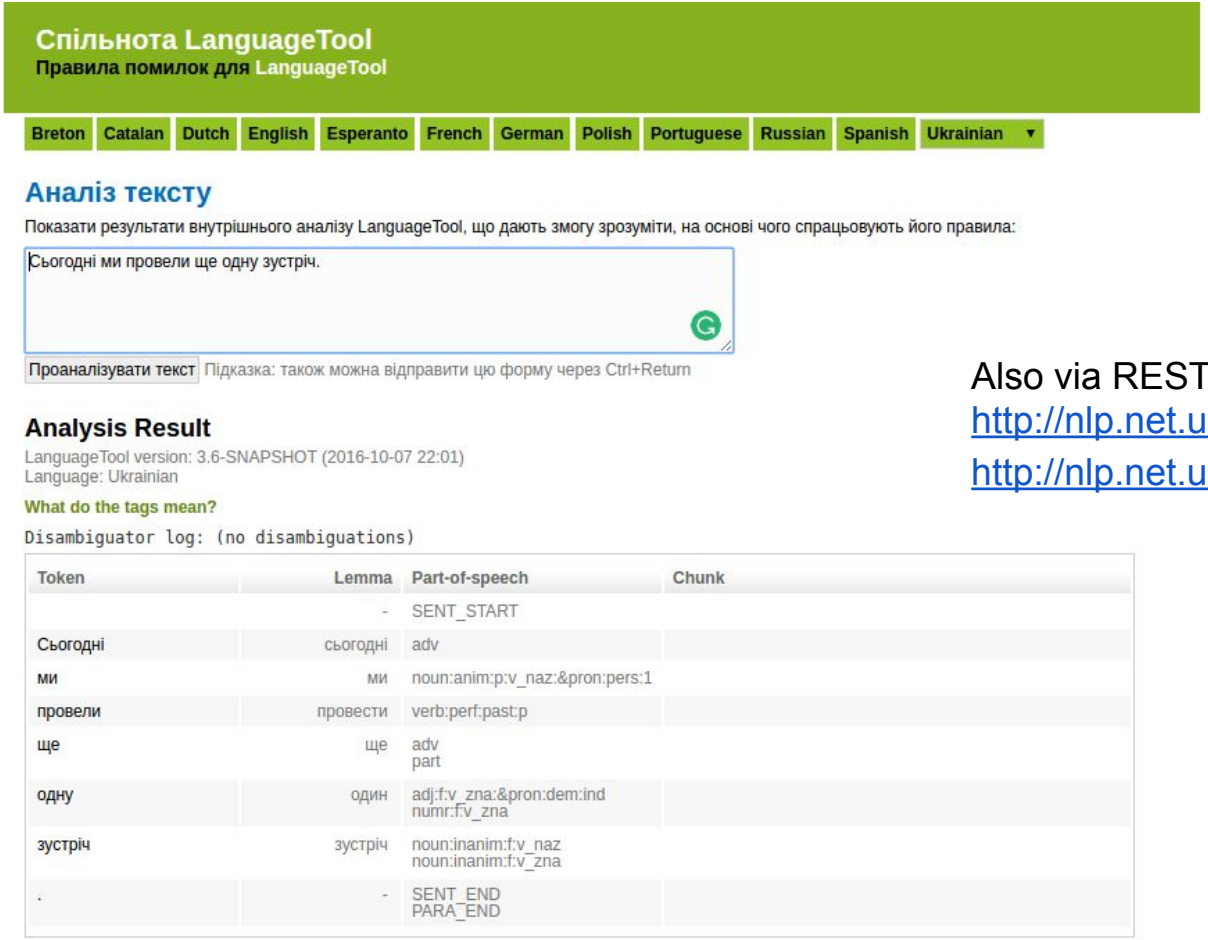
Linguist, PhD

... and many others

LanguageTool for Ukrainian languagetool.org/uk



http://community.languagetool.org/analysis/analyzeText



Also via REST API

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

http://nlp.net.ua:8787/dict_uk/swagger-ui.html

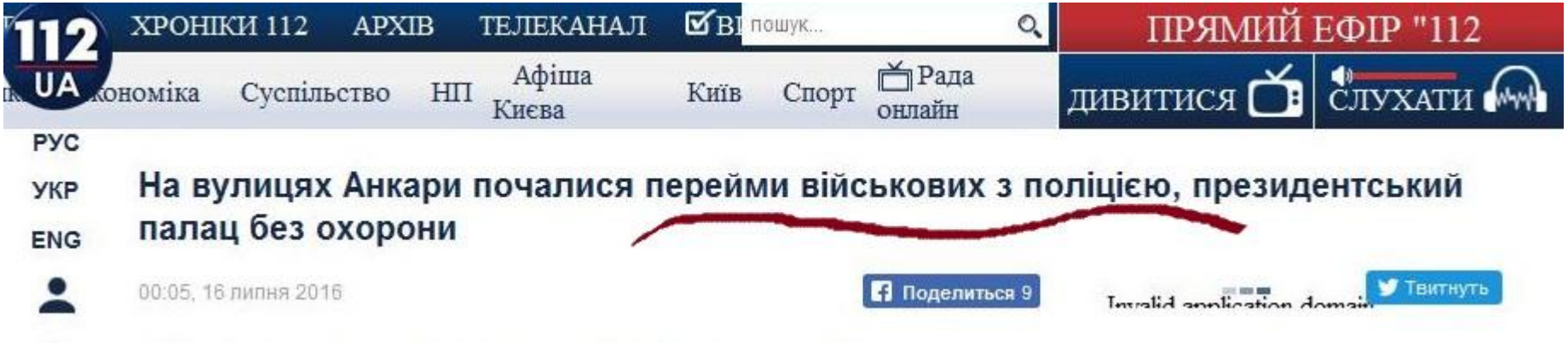
Latest version of LanguageTool 3.5 for Ukrainian

- dictionary with over 208,000 lemmas

  ВЕСУМ: 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 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 considerinmg

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)

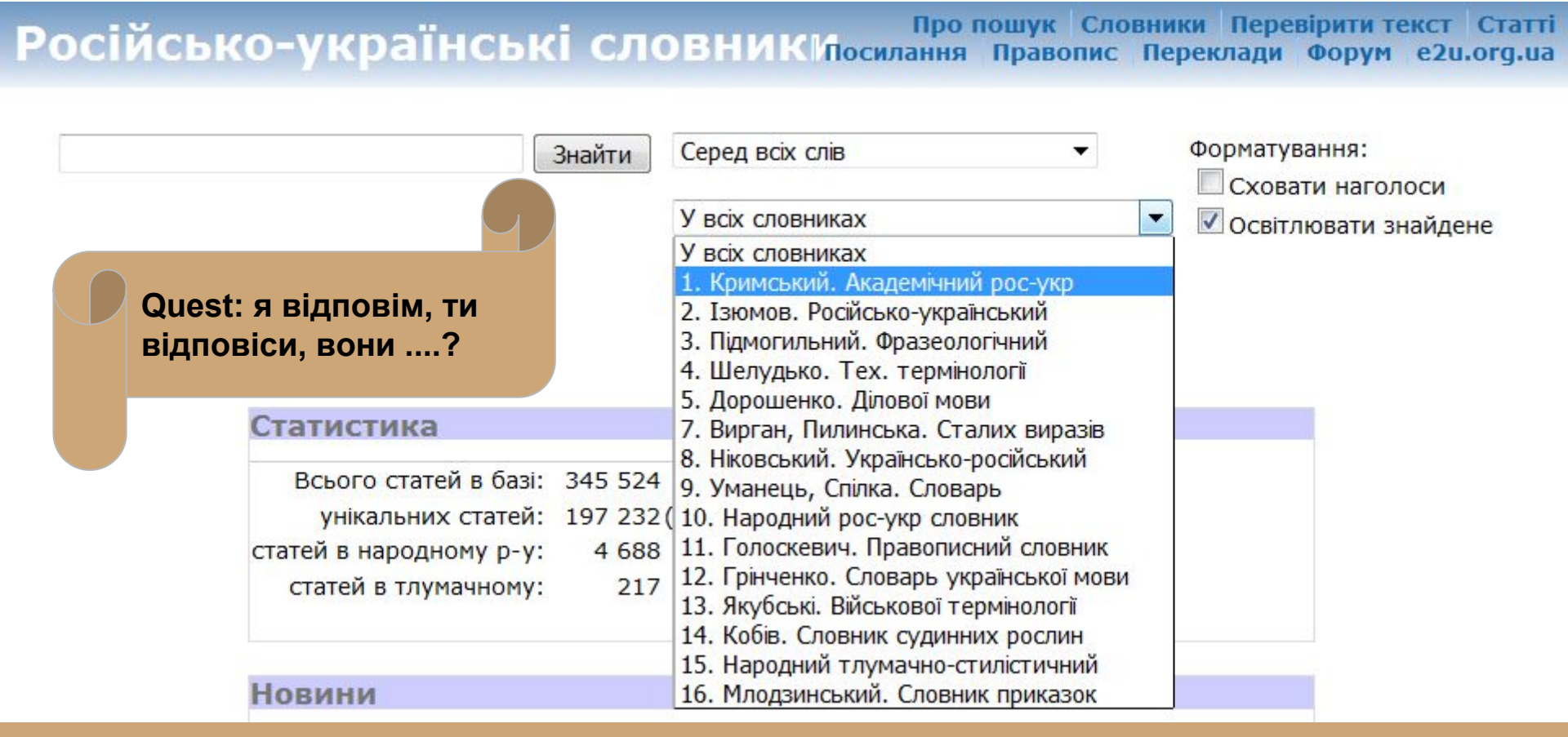
"приналежність" замість "належність"

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

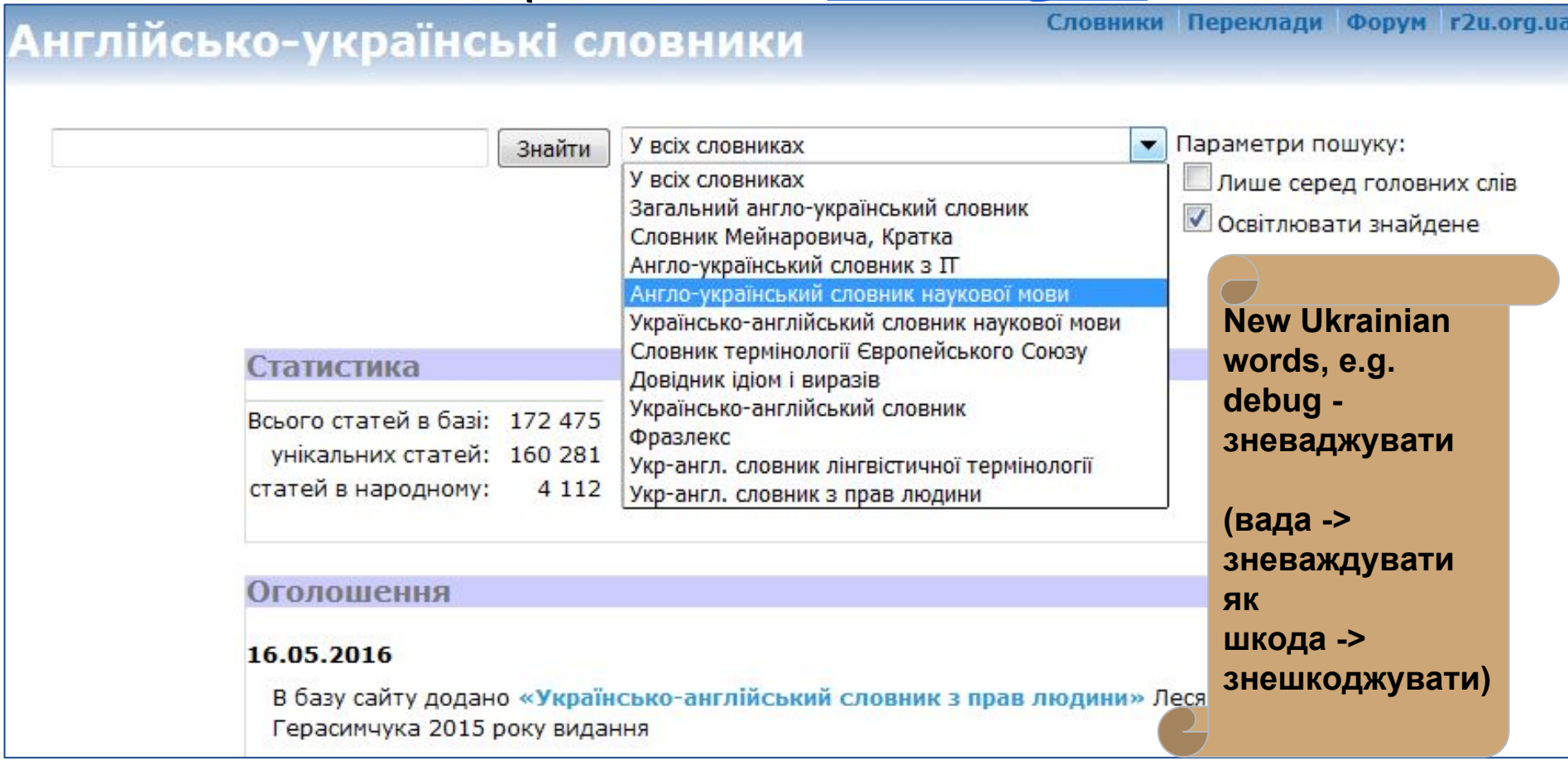
Next steps

- 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



Keep in mind: e2u.org.ua



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

GitHub: Brown-uk

nlp_uk

dict_uk

corpus

Find more about us: http://r2u.org.ua/corpus
Facebook: www.facebook.com/r2u.org.ua/

Thank you, you are the best!

Any questions?