LREC-COLING 2024

BRUK Team's Resources for Ukrainian Corpus Creation

BRUK and USL

Vasyl Starko, Ukrainian Catholic University

UNLP 2024

Ukrainian Brown Corpus (BRUK)

- Loosely modeled on the original Brown Corpus
- One million words
- Stratified random sampling, samples of up to 2,000 words
- Genre-balanced, POS-tagged
- Focus on edited written Ukrainian (original, human-written, high quality)
- Timeframe: 2010–2020
- Detailed metadata stored separately
- Goal: create a POS gold standard of Standard Modern Ukrainian
- More info: aclanthology.org/2023.unlp-1.11/

Text categories in BRUK

A. Press, 25%

B. Religion, 3%

C. Skills and hobbies, 7%

D. Essays, biography, memoirs, etc., 7%

E. Administrative documents, 3%

F. Popular science, 5%

G. Science, 10%

H. Textbooks, 15%

I. Fiction, 25%

Disambiguation

Over 700,000 tokens (and counting) fully disambiguated

```
<sentence>
  <token value="Як" lemma="як" tags="conj:subord" />
  <token value="дисидент" lemma="дисидент" tags="noun:anim:m:v_naz" />
  <token value="я" lemma="я" tags="noun:anim:s:v_naz:&amp;pron:pers:1" />
  <token value="народився" lemma="народитися" tags="verb:rev:perf:past:m" />
  <token value="у" lemma="у" tags="prep" />
  <token value="кабінеті" lemma="кабінет" tags="noun:inanim:m:v_mis" />
  <token value="офіцера" lemma="офіцер" tags="noun:anim:m:v_rod" />
  <token value="КГБ" lemma="КГБ" tags="noun:inanim:m:v_rod:nv:abbr:prop:bad" />
```

BRUK: Conclusions and future work

- (projected) POS gold standard for Modern Ukrainian
- 700+k tokens fully disambiguated
- Available for download and training language models

github.com/brown-uk/corpus

- Supplement BRUK with other types of texts
  Add a semantic annotation layer

Ukrainian Semantic Lexicon (USL)

Motivation:

- Need for semantic annotation in Ukrainian corpora (GRAC, BRUK, etc.)
- No suitable publicly available resource

Goal: Create a semantic resource that is

- integrated with other available tools for Ukrainian
- machine-readable, downloadable
- expendable and customizable

USL development was supported by a grant from the Believe in Yourself Foundation at UCU

USL: key principles

- Faceted (not hierarchical) approach to classification
- Linguistically meaningful taxonomic classes
- Coarse-grained sense division (98% of USL 2.0 entries have 1 sense)
- Integration with VESUM and TagText

For more details see:

V. Starko, Implementing Semantic Annotation in a Ukrainian Corpus. Computational Linguistics and Intelligent Systems. Proc. 5th Int. Conf. COLINS. Vol. I. (2021): 435–447

V. Starko, Semantic Annotation for Ukrainian: Categorization Scheme, Principles, and Tools. Computational Linguistics and Intelligent Systems. Proc. 4th Int. Conf. COLINS. Vol. I. (2020): 239–248

Examples of USL entries

noun

науковець ‘scholar’ 1:conc:hum:prof

adjective

великий ‘large, great’ 1:size:2:degree:3:age

adverb

повністю ‘completely’ 1:physqual:2:degree:max

verb

стояти ‘to stand’ 1:loc:body:noncaus:2:loc:noncaus

Multifaceted tag assignment

& — simultaneous semantic features:

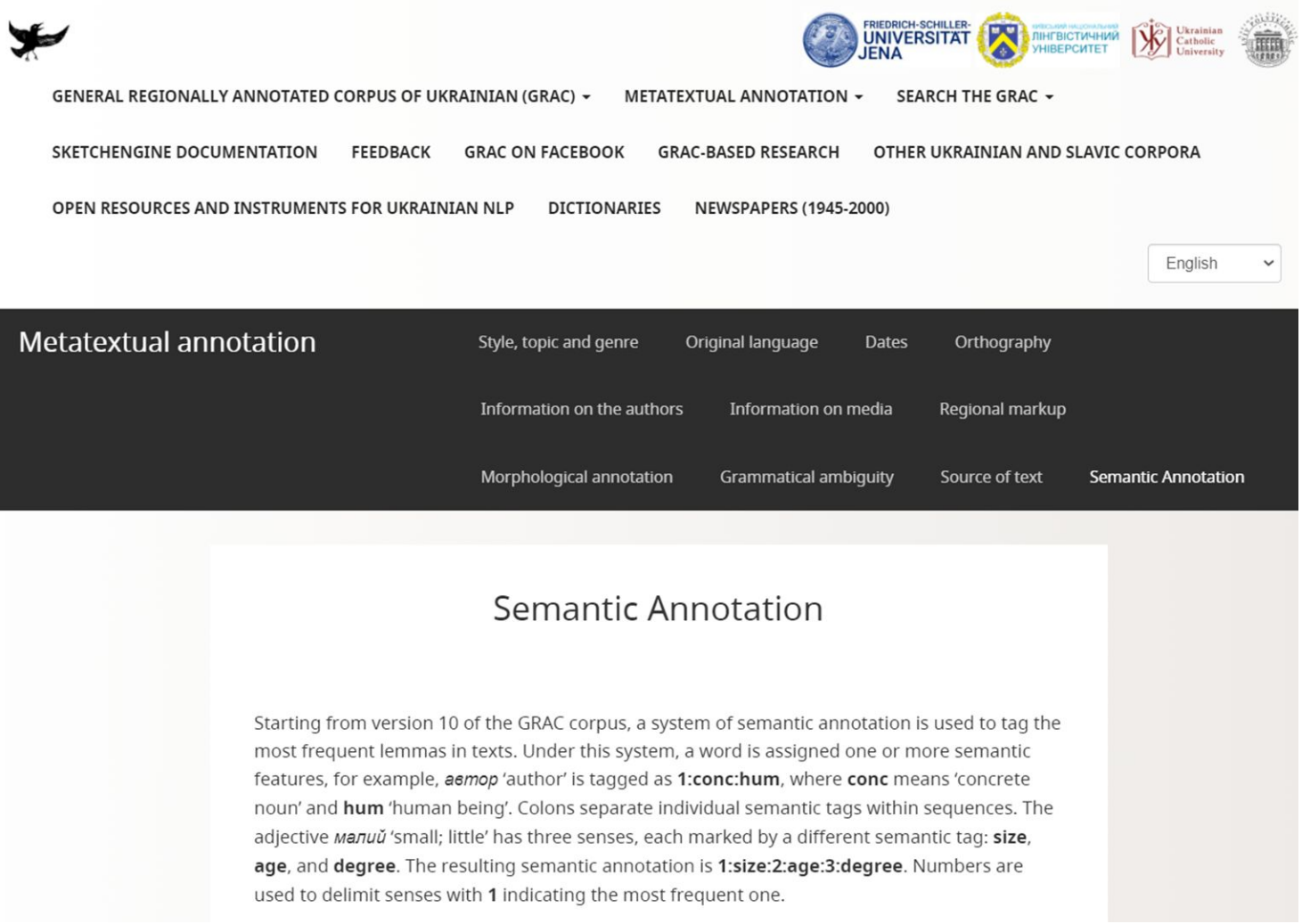
abst:time:period&unit is assigned to

хвилина ‘minute’

година ‘hour’

день ‘day’

etc.



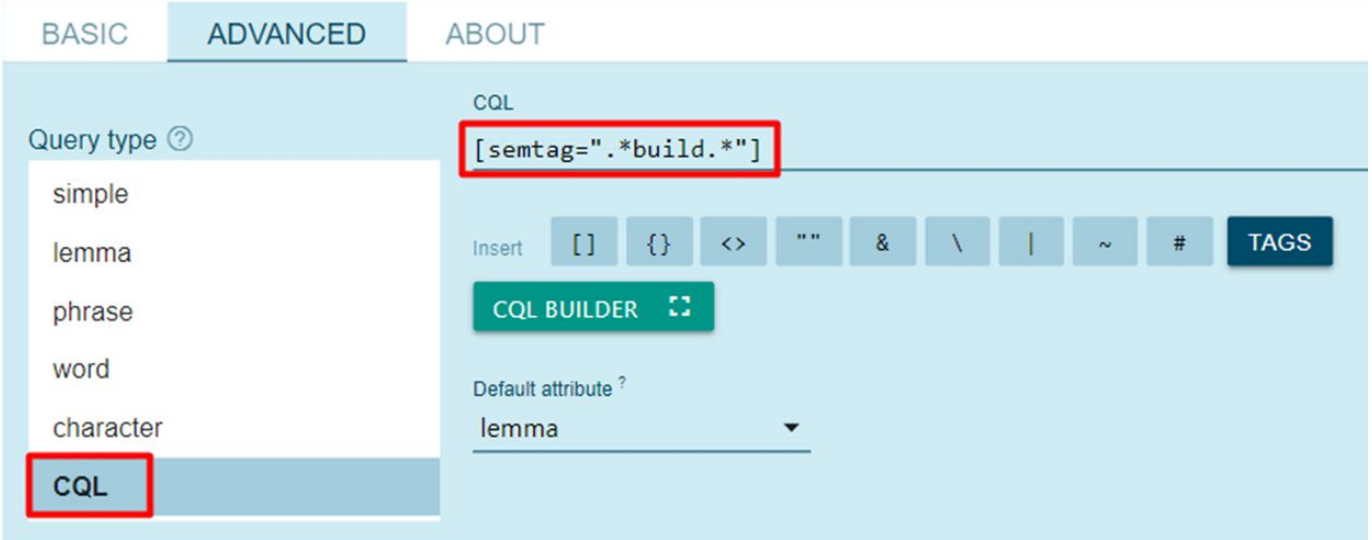
Full semantic dataset at

uacorpus.org

Lexical coverage

- **46%** (GRAC, USL 2.0)
- Some high-frequency classes (pronouns, conjunctions, and particles) that are outside of the scope of semantic annotation account for around 30% of words in the corpus
- Theoretical ceiling for semantic coverage — 70%

Semantic query in GRAC



The query matches all nouns with the tag build anywhere in the sequence:

conc:build buildings and constructions

conc:org&&build organizations and buildings

conc:build:part parts of buildings

Semantic tagging

- USL is available as part of VESUM at github.com/brown-uk/dict_uk
- TagText can perform morphosemantic tagging in one pass

```
<sentence>
  <tokenReading>
    <token value='Сьогодні' lemma='сьогодні' tags='adv' semtags='1:time:deictic' />
  </tokenReading>
  <tokenReading>
    <token value='відбулася' lemma='відбутися' tags='verb:rev:perf:past:f' semtags='1:exist
  </tokenReading>
  <tokenReading>
    <token value='важлива' lemma='важливий' tags='adj:f:v_kly:compb' semtags='1:abst' />
    <token value='важлива' lemma='важливий' tags='adj:f:v_naz:compb' semtags='1:abst' />
  </tokenReading>
  <tokenReading>
    <token value='подія' lemma='подія' tags='noun:inanim:f:v_naz' semtags='1:abst:event' />
  </tokenReading>
```

USL: Conclusions and future work

- Ukrainian Semantic Lexicon (USL 2.0) has 81,000+ lemmas
- USL and its full semantic tagset are available online
- Semantic annotation in USL is faceted and coarse-grained
- GRAC has been annotated with USL 2.0
- Next: expand USL and modify the tagset as needed

LREC-COLING 2024

BRUK Team's Resources for Ukrainian Corpus Creation

Vasyl Starko, Ukrainian Catholic University (Ukraine)
Andriy Rysin, Independent Researcher (USA)

Andriy Rysin

- SAS Institute, NC, USA
- Open-source enthusiast
- Supporting Ukrainian in IT for over 25 years
- arysin@gmail.com

VESUM
https://github.com/brown-uk/dict_uk

- History: ispell-uk (199x) -> aspell-uk -> hunspell-uk
  - hunspell used by: Firefox, LibreOffice/OpenOffice, PostgreSQL, other open-source software
  - Derivative dictionary: Google Chrome
- Spellchecking:
  - simple set of unique word forms
- Grammar checking:
  - The need to tell cases, gender etc
  - hunspell-uk -> VESUM
- POS Tagging
  - Corpora: GRAC, BRUK…

VESUM

More than just a morphological dictionary:

- Over 420000 lemmas
- The full version is compiled from the compact source files
- Flexible system to specify morphological groups (e.g. голова /n10.p)
- Contains non-standard categories: bad, slang, vulg, alt, arch…
- Contains non-standard forms: long, short, subst…
- Contains additional tags: compb/c/s, adjp, geo, namef/l/p…
- Contains tags for spelling norms of 1992 and 2019
- Pronouns belong to other groups and have an extended system of pronoun tags (by Vyhovanets, Horodenska)
- Contains case government information (for prepositions, verbs etc)
- Contains replacements for over 8000 покручів
- Credits: ...

VESUM - Additional Forms

- :long - красную, терпіте, стріляючися
- :short - красен, співа, допомагать, гарніш
- :bad - автомобілей, альманаха, болгарів
- :v_zna:var - (побачив) стола
- :v_zna:rare - (пішов у) президенти
- :subst - робе, їсиш

VESUM - Additional Lemmas

- од- - одповісти
- alt - аксіяльний, амплюа, февдалізм
- slang - актівіті, башляти, зашибісь
- vulg - срака
- bad:
    - :adjp:actv:imperf - існуючий
    - -учи - будучи, хочучи
    - Russianism - кляча
    - frequent spelling mistakes - біржевий, грунт, експресо
    - шо, шоби, благодіяти, відціля, сьодні

VESUM - Other specifics

- Abbreviations with/without a dot:
  - РАЦС, кг, смт, р., с.
- Large number of proper nouns
  - names, surnames, patronyms, toponyms, companies etc
- Lemmas and forms present only in either spelling 2019 or spelling 1992:
  - проєкт vs проект
  - Мюнхену (р.в.)
  - артбомонд vs арт-бомонд

VESUM - Other specifics

- Originally was cross-referenced with existing dictionaries
- Quickly went beyond any existing dictionaries
- For now the main source is Ukrainian corpora
- Priority - practical applications
  - Orthography
  - Grammar
  - Indexing
  - Analysis
- Practical use gives rapid feedback

Development

- General challenges:
  - Dual gender - бабище, вовчище, вітрище
  - Genitive case, 2 declension, masc.
    - Form separation: смарагд (мінерал: -у, камінь: -а)
  - пекло - «пекел» - frequent in GRAC (ГС - пекл, УЛІФ - only singular)
  - привид: anim / inanim (ГС, УЛІФ - only inanim, gen:. -у)
  - Plural for last names - ч.р. чи ж.р.?
  - Genitive case for plural for animals: пасти корови…
  - Complicated homonyms for substantiated adjectives: безпартійний, поранений, майбутнє, українська
  - Multi-gender non-inflected nouns: ауді, агреже, есемес
  - Vocative for inanim: «форуме» - 0 occurrences in Ukrainian words, but present as Russian one
- New words:
  - Words from GRAC, BRUK, and modern texts
  - Frequency, multiple authors…
  - Figuring out the proper tags (arch, alt, bad, slang etc)
  - Genitive case, 2 declension, masc. Ending (-а/-у)
  - Unnecessary homonyms
  - Alternative spelling - sometimes we’ll accept many
- Practicality over theoretical “correctness”
- Impossible to cover all combined words - LT dynamic tagging comes to the rescue

Indexing: Lucene/ElasticSearch

- Ukrainian modules in Lucene (full-text search system) / ElasticSearch
  - Uses lemmatizer instead of a stemmer
  - Used for search in Ukrainian Wikipedia

LanguageTool

- https://languagetool.org/
- Supports 25 languages
- Ukrainian module
  - Based on VESUM
  - Over 1000 rules for grammar, style, and punctuation checking
    - Token agreement: prep+noun, adj+noun, noun+verb, verb+noun (VESUM has case government info)
  - Dozens of rules for simple disambiguation
- Analysis steps:
  - Sentence tokenization (srx)
  - Word tokenization
  - Tagging
    - Dynamic tagging: 11-й, по-зимбабвійські, Порошенко-старший…
    - Additional tags: number, number:latin, date, time, hashtag…
    - Alternative spelling: ґубернія, термометер…
  - Rule-based disambiguation (xml, Java)
  - Checking rules (xml, Java)
  - Other functionalities
- Developer's webpage: https://dev.languagetool.org/

All changes (VESUM, tagging, rules) are tested for regression on a 500M-word modern texts

- Support for Crimean Tatar language was added in 2024

NLP tools for Ukrainian

- NLP_UK (https://github.com/brown-uk/nlp_uk)

  Set of tools for analysis of Ukrainian texts

- TagText
  - Utility for POS-tagging / semantic tagging of Ukrainian texts
  - Is based on LanguageTool core
  - Provides additional tags: punct, symb, unknown, unclass…
  - Allows to generate results in xml, json, or text formats
  - Has flexible parameters to control the tagging
  - Allows to collect statistics
- Used by:
  - GRAC
  - BRUK
  - UberText 2.0
  - …

NLP tools for Ukrainian

Statistics-based disambiguation

- It can disambiguate based on statistics from BRUK
- Initially was designed to help with tagging BRUK corpus
- For modern texts gives fairly high precision:
  - 99% for lemmas
  - 98% for lemma/POS
  - 91-95% for a full (VESUM) tagset
    - ~30% of errors are nom vs acc (v_naz/v_zna)
- Statistics is built on 700k tokens from BRUK (test set: 50k)
- Shows very stable result starting with a training set of 200k tokens
- Uses left/right token for the context (3 attributes: token, lemma, POS-tag)
- Works in real-time; no special memory/CPU requirements
- Could be improved with neural network instead of a fixed logic

CleanText

Most texts that arrive for analysis have issues (OCR problems, typos, markup processing, word-wrapping etc.)

CleanText helps with problems like these:

- Broken/mixed encodings (cp1251 etc)
- Cyrillic/Latin mix (л**i**тера)
- Numbers instead of letters («3а» with number 3 instead of letter 3)
- Normalization problems: e.g. «ї» and «й» with combined characters
- Non-standard apostrophes
- Removes soft hyphen (U+00AD) and other hidden symbols
- Removes word wrapping (based on the dictionary)
- Can mark/cut paragraphs in Russian
- Cannot deal with letter spacing: Н а т а л к а і в о в к

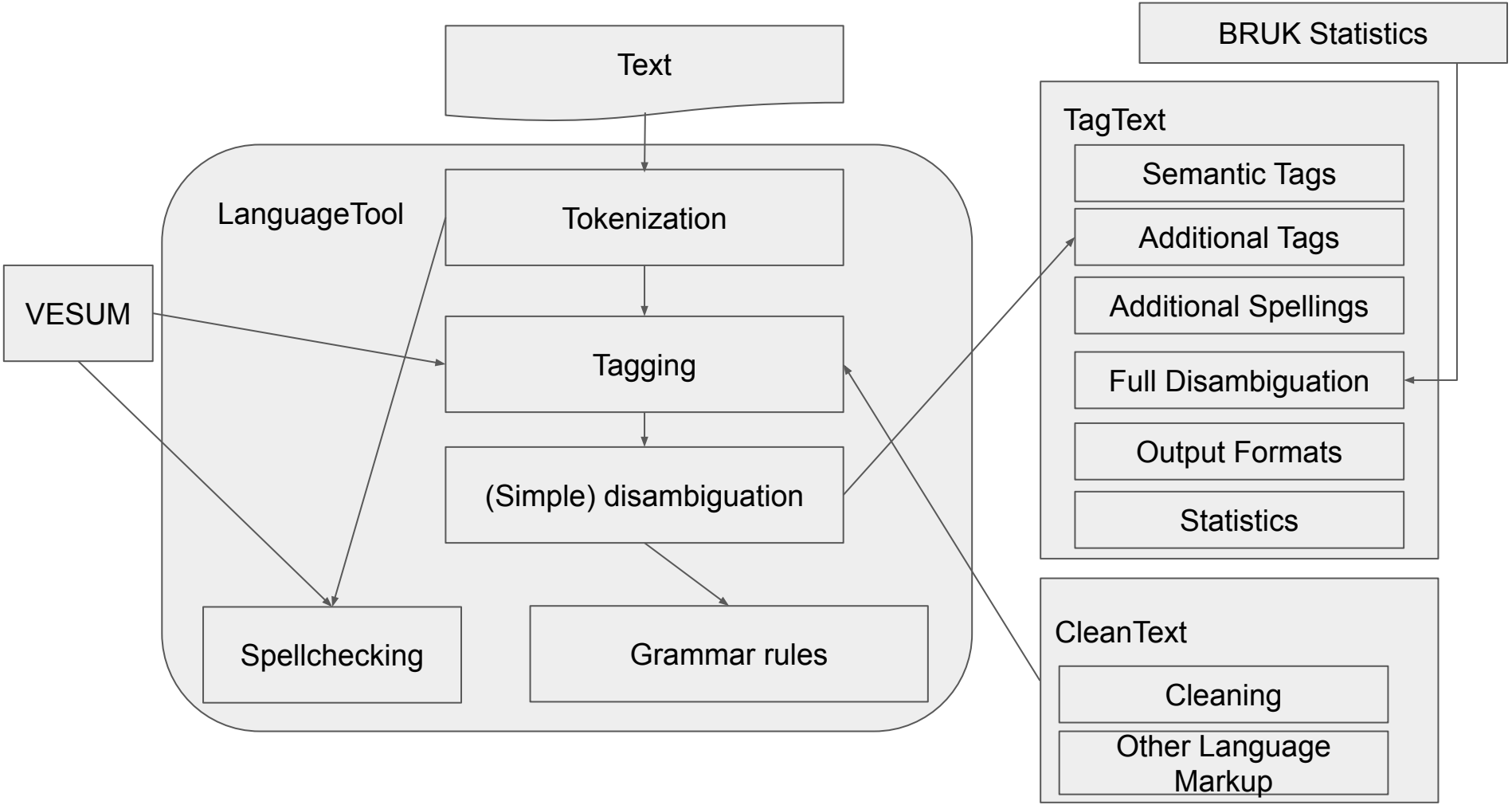
GRAC Corpus

General Regionally Annotated Corpus of Ukrainian

- Almost 2 billion tokens; many regions and a wide date range
- Has non-standard spelling systems: Zhelehivka, mails from L. Ukrainka…
  - Has large number of archaic, dialectic, and slang words
- Uses tagset from VESUM/nlp_uk
- Uses CleanText + TagText
- Currently TagText recognizes and tags ~99% of the words in the texts with modern spelling
- Plans to use statistical disambiguation

BRUK (БрУК) Corpus

- https://github.com/brown-uk/corpus
- Built on the principles of Brown Corpus
- Uses tag set from VESUM/nlp_uk
- The final result is proofread
- Results are verified using LanguageTool grammar rules
  - Token agreement rules
- Currently 705k tokens are fully disambiguated (test set: 50k)


Text
BRUK Statistics
TagText
Semantic Tags
Additional Tags
Additional Spellings
Full Disambiguation
Output Formats
Statistics
LanguageTool
Tokenization
VESUM
Tagging
(Simple) disambiguation
Spellchecking
Grammar rules
CleanText
Cleaning
Other Language Markup

We Appreciate Feedback

- All projects are open source
- Free to use
- Feedback is appreciated
  - Page on FB: https://www.facebook.com/pravopysnyk.lt
  - Issues on https://github.com
  - Forum: https://r2u.org.ua/forum