Ural Federal University
named after the first President of Russia B.N.Yeltsin

Yet Another RussNet: Spinning-in-Progress

P. Braslavski, D. Ustalov, M. Mukhin, Y. Kiselev
Ural Federal University
Yekaterinburg, Russia

Outline

- Introduction
- Related Work
- Structure
- Implementation
- Evaluation
- Conclusion

Introduction

- A thesaurus is a critical resource for successful NLP and AI applications.
  - No open source thesaurus for Russian.
- **Yet Another RussNet**, started in 2013, is aimed at creation of such one.
  - Crowdsourcing is used.

http://russianword.net/en/

Related Work: Thesauri

- Created by **experts**:
  - RussNet, RuThes(-lite), UNL.
- **Crowdsourcing**:
  - Russian Wiktionary.
- **Automatically** derived:
  - BabelNet, WordNet.ru, Russian Wordnet.
- Other **Slavic** languages:
  - plWordNet, BulNet, etc.

Related Word: Crowdsourcing

- Three genres of crowdsourcing.
- **Games with a purpose**.
  - Ex.: Phrases Detectives, JeuxDeMots, etc.
- **Mechanized labor**.
  - Ex.: Mechanical Turk, CrowdFlower, etc.
- **Wisdom of the crowds**.
  - Ex.: Wikipedia, Wiktionary, etc.

YARN Structure

- **YARN** is conceptually similar to **PWN**.
- Words have grammatical features.
- Synsets may contain glosses.
- Words in synsets can have definitions, usage examples, and labels.
- Each synset may belong to a domain.

YARN Structure: Relations

- Synsets are linked to each other primarily via *is-a* relations.
- Other relations:
  - meronymy (*part-of*) between synsets,
  - antomymy between lexical senses.
- We elaborated 4-5 top levels for each part of speech.

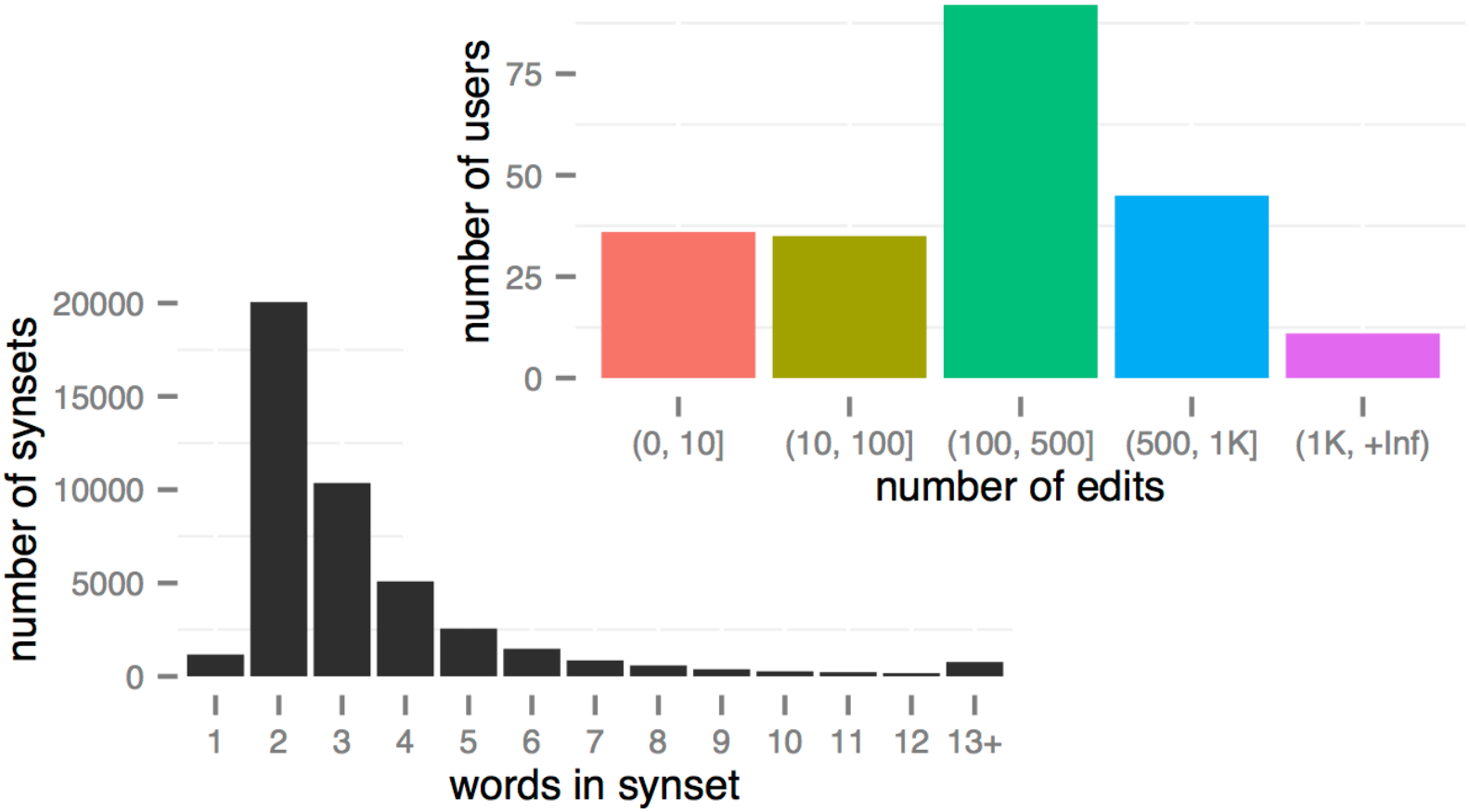
YARN Structure: Raw Data

- YARN is not created from scratch.
- The following “raw data” are used:
  - Wiktionary (as the core),
  - Wikipedia redirects,
  - UNLDC,
  - Russian National Corpus statistics.
- The goal of YARN is to refine these data to make a successful resource.

Current State of YARN

- **More than 200 people** have taken part in the synset assembly.
- The resource comprises **100K+ words** and **46K+ synsets** under **CC BY-SA**.
- WotC: https://russianword.net/editor.

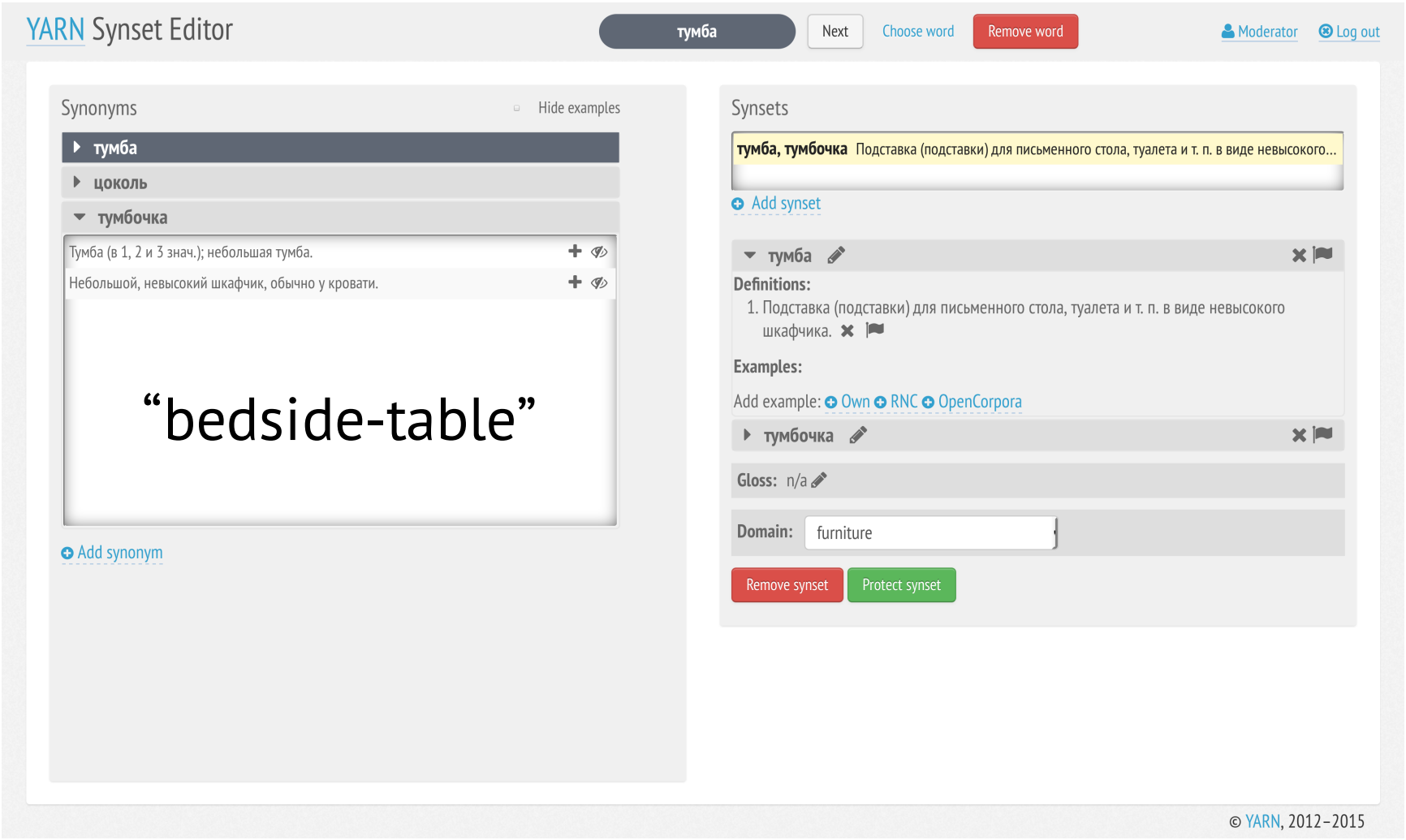
Synsets & Users



Implementation Details

- Web-based app: **Ruby on Rails**.
- Data are internally stored in a **PostgreSQL** database.
  - Aggregated action chunks are tracked.
- Export formats: XML, Turtle, CSV.
- Import formats: XML, CSV.
- Schema is (somewhat) similar to LMF.

User Interface



Synset Definition

```
<synsetEntry id="s9439" author="122" version="29" timestamp="2014-11-17T07:49:46Z">
  <word ref="w9244">
    <definition source="ru.wiktionary" url="http://ru.wiktionary.org/wiki/суп">
      Жидкое кушанье, обычно представляющее собой отвар с приправами и употребляемое как
    </definition>
    <example source="'Путешествие в седьмую сторону света', 2000, НКРЯ.">
      Был обед - овощной суп и курица на второе.
    </example>
  </word>
  <word ref="w40078"/>
  <word ref="w2893"/>
</synsetEntry>
```

{суп, бульон, похлёбка (*soup*)}

Current Problems

- **Organizational issues**.
  - The number of synsets was growing; moderators were not able to assess edits.
- **Synset duplication**.
  - Participants do not consult other people's work.
- **Hyponymy confusion**.
  - In some cases, participants mix hyponymy with synonymy.

Genre → mechanized labor?

Evaluation: Size

- We compared YARN with other Russian thesauri.

Table 1: Russian thesauri comparison.

	# of concepts	# of relations	# of words	Availability	Commercial Usage
RussNet	5.5K	8K	15K	No	No
Russian Wordnet	157K	—	124K	No	No
RuThes	55K	210K	158K	No	No
RuThes-lite	26K	108K	115K	Yes	No
YARN	44K	0	48.6K	Yes	Yes

The table is present in the paper.

Evaluation: Quality

- We took **200** most frequently edited synsets and assessed the quality of each synset.
  - Scale: Excellent, Satisfactory, Bad.
- Then, we aggregated the 800 answers using the MV strategy with pessimistic ties.

Evaluation: Quality

- Krippendorff's alpha $\alpha = 0.202$ due to the skewness of the answers.
- Given these results, we treat the top 200 synsets as sufficiently good.

Table 2: YARN synset quality.

	MV	1	2	3
Excellent	103	37	62	21
Satisfactory	70	3	43	11
Bad	27	0	12	11
Total	200	40	117	43

Conclusion

- The deliverables of **YARN** are available on its website under **CC BY-SA** in **XML**, **CSV**, **RDF.**
- https://russianword.net/en/
- https://nlpub.ru/YARN
- https://github.com/russianwordnet

Future Plans

- Creating verb and adjective synsets.
- Establishing more relations.
- Development of automatic methods for quality assurance.
- Widening the audience.
- Development of crowd management techniques.

Thanks!

Dmitry Ustalov,

Ural Federal University.

- https://ustalov.name/en/
- dmitry.ustalov@urfu.ru

The present work is supported by the Russian Foundation for the Humanities, project № 13-04-12020, and by the Mikhail Prokhorov Foundation.