UraLUID: Supporting data-driven (prosodic) research

Eszter Simon, Nikolett Mus, Ágnes Kalivoda, Eszter Ruttkay-Miklián

Research Institute for Linguistics, Hungarian Academy of Sciences

{*simon.eszter,mus.mikolett,kalivoda.agnes*}*@nytud.mta.hu, eruttkaym@gmail.com*

2nd workshop on Uralic prosody
September 28-29, 2017, Budapest

Outline

Background

- *Languages under influence. Uralic syntax changing in an asymmetrical contact situation*
- Research Institute for Linguistics, Hungarian Academy of Sciences
- January 2016 – July 2017
- supported by the National Research Development and Innovation Office
- website: `http://www.nytud.hu/depts/tlp/uralic/dbases.html`

Languages under influence

The pilot-project had a twofold objective:

- Theoretical work
  - to describe and analyze **potential syntactic** changes due to Russian contact
  - examined languages: **Udmurt**, **Northern Khanty** (Synya dialect), **Eastern Khanty** (Surgut dialect group), and **Tundra Nenets**
- Practical work
  - to collect and structure **primary data**
  - to build a linguistically annotated database of **written** and **spoken** texts, as well as of **"old"** and **"new"** data

Source types

	Udmurt	Synya Khanty	Tundra Nenets	Surgut Khanty
“old”	folklore	folklore	folklore	folklore
“new”	blog post	interview	newspaper spontaneous speech folklore	interview spontaneous speech

Data types

	Udmurt	Synya Khanty	Tundra Nenets	Surgut Khanty
“old”	written	written	written	written spoken
“new”	written	spoken	written spoken	written spoken

Spoken data

	Speakers	Place of recording	Length	Format
Synya Kh.	1	Ovgort	01:08:22	`.wma`
Tundra Nenets	1	Moscow	01:57:40	`.wav`
Surgut Kh. (new)	2	Kogalym	00:28:58	`.wav`
Surgut Kh. (old)			00:36:17	`.wav`

The Synya Khanty spoken data

- the original data were collected within the frame of the project entitled **In the Khanty way** (`http://hantisirn.nytud.hu/`)
- source: interviews
- genre: expository, narrative texts

The Tundra Nenets and Surgut Khanty spoken data

- the spoken data were collected during our fieldworks
- source: spontaneous speech data (elicited by graphical materials), "read" speech
- genre: narrative, procedural, expository texts

Data available on our website

1. Metadata
   - title of the text
   - text ID
   - pages
   - file name
   - genre
   - token number
   - (sub)dialect
   - age of the speaker
   - gender of the speaker

Data available on our website (cont.)

`http://www.nytud.hu/depts/tlp/uralic/dbases.html`

2. Text
   - original transcription
   - Cyrillic
   - FU transcription(s)
   - **IPA**

   $\Rightarrow$ the original text split into sentences

Data available on our website (cont.)

3. Morphologically analyzed text: `.tsv` files consisting of 15 columns
   - 1-9: the token and its transcriptions
   - 10: (segmented) token
   - 11: lemma
   - 12: Hungarian gloss
   - 13: English gloss
   - 14: POS tag (and semantic label if there is any)
   - 15: RUS label (if the word has Russian origin)

Data available on our website (cont.)

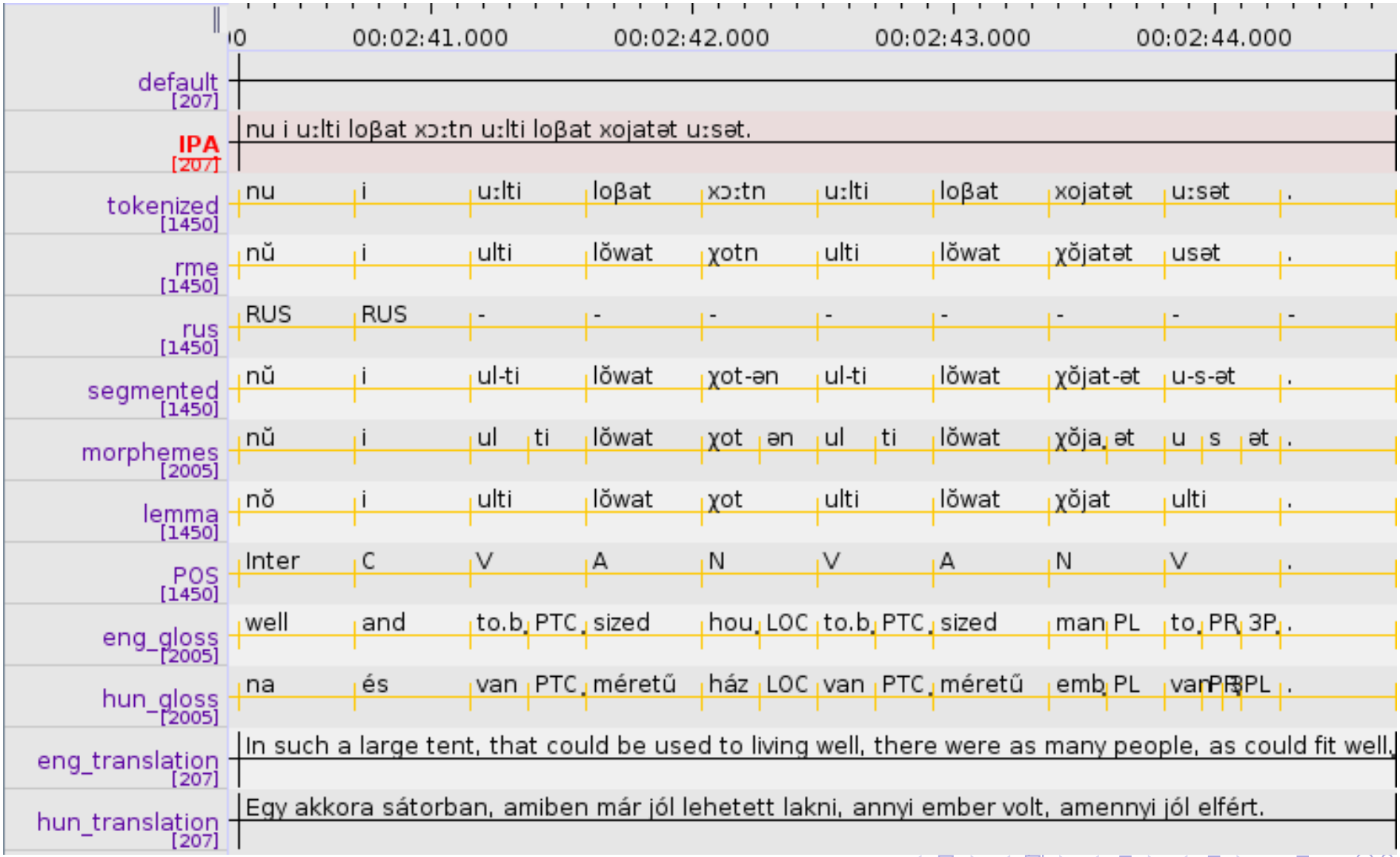
4. Translation
   - **English**
   - Russian
   - German
   - Hungarian

   $\Rightarrow$ every translation is sentence-level aligned with the original text

Data available on our website (cont.)

5. ELAN files
   - `.eaf` file (containing data on sentence-, token- and morpheme-level)
   - the `.eaf` files are created with a script written in Python3
   - the script uses the Pympi module developed for creating and processing ELAN and Praat annotation files (`http://github.com/dopefishh/pympi`)
   - the original sentences are aligned to the time slots of the audio file
   - the other pieces of information are connected to the sentences via symbolic references

Data in ELAN


00:02:41.000
00:02:42.000
00:02:43.000
00:02:44.000
default [207]
IPA [207]
nu i u:lti loβat xɔ:tn u:lti loβat xojatət u:sət.
tokenized [1450]
nu i u:lti loβat xɔ:tn u:lti loβat xojatət u:sət .
rme [1450]
nŭ i ulti lŏwat χotn ulti lŏwat χŏjatət usət .
rus [1450]
RUS RUS - - - - - - - -
segmented [1450]
nŭ i ul-ti lŏwat χot-ən ul-ti lŏwat χŏjat-ət u-s-ət .
morphemes [2005]
nŭ i ul ti lŏwat χot ən ul ti lŏwat χŏja ət u s ət .
lemma [1450]
nŏ i ulti lŏwat χot ulti lŏwat χŏjat ulti .
POS [1450]
Inter C V A N V A N V .
eng_gloss [2005]
well and to.b PTC sized hou LOC to.b PTC sized man PL to PR 3P .
hun_gloss [2005]
na és van PTC méretű ház LOC van PTC méretű emb PL van PR 3PL .
eng_translation [207]
In such a large tent, that could be used to living well, there were as many people, as could fit well.
hun_translation [207]
Egy akkora sátorban, amiben már jól lehetett lakni, annyi ember volt, amennyi jól elfért.

How to use the UraLUID database?

1. use the `.eaf` files with the corresponding audio files:
   - download the latest version of ELAN: `https://tla.mpi.nl/tools/tla-tools/elan/`
   - download the Charis SIL font package: `https://software.sil.org/charis/`
   - download the `.eaf` and audio files: `http://www.nytud.hu/depts/tlp/uralic/dbases.html`
   - open ELAN → Open... → choose the needed `.eaf` file and the corresponding audio file
2. use the `.tsv` files
   - download the `.tsv` files
   - use Unix commands or your dear old statistical tools

Future plans

Tundra Nenets spoken data:

- Transcription
  - Cyrillic
  - FUT
  - IPA
- Morphological analysis $\Rightarrow$ improving the Giellatekno's morphological analyzer
- POS-tags
- Translation
  - Russian
  - English

Acknowledgements

The authors wish to thank Elena Skribnik, Zsófia Schön, Agrafena Pesikova, Anisya Volkova, Katalin Gugán, Márta Csepregi, Olesya Khanina, Aleksey Kozlov, Khadry Okotetto, Orsolya Tánczos, Erika Asztalos, Yulia Speshilova, and Szilvia Németh for their contribution to the building of the database.

Thank you for your attention!

`http://www.nytud.hu/depts/tlp/uralic/dbases.html`