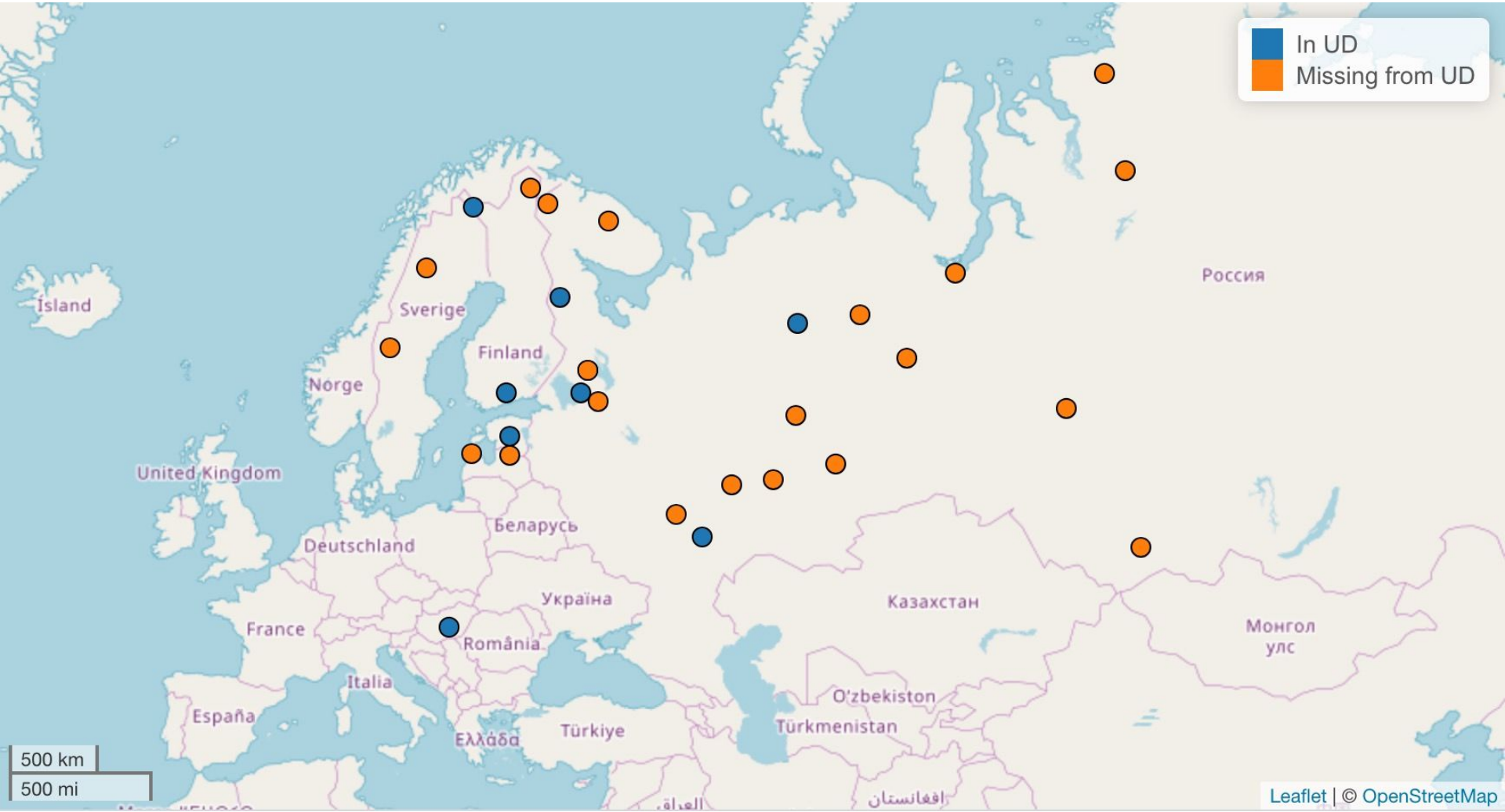
Survey of Uralic Universal Dependencies development

Niko Partanen & Jack Rueter

University of Helsinki

Uralic languages

- A large language family in Northern Eurasia
  - Approximately 38 languages
- Regular morpho-semantic complexity
- Relatively free constituent ordering
- Both closely and distantly related languages


In UD
Missing from UD
Россия
Ísland
Sverige
Finland
Norge
United Kingdom
Беларусь
Deutschland
Україна
France
România
Italia
España
Ελλάδα
Türkiye
Казахстан
Монгол улс
O'zbekiston
Türkmenistan
افغانستان
العراق
500 km
500 mi
Leaflet | © OpenStreetMap

Uralic treebanks – current status

- 11 treebanks in 7 Uralic languages
- Missing major branches: Mari, Ob-Ugric and Samoyedic
- Geographically Siberia still a missing area
- Largest languages best represented

Uralic treebanks – assumptions

- As all treebanks are annotated with the same system, it would be reasonable to expect that especially closely related languages are annotated similarly
- Some differences are to be expected – these are still different languages

- Differences possible at all levels:
    - Lemmatization
    - Morphological tags
    - Dependencies used

Consistency??

- Maximal comparability between treebanks would be **desirable**
- Since the languages are related and not entirely dissimilar, having consistent annotations should be easier to achieve than between unrelated languages

- There will be **new Uralic treebanks**, a common ground on annotations would make initiating this work easier

Example: Personal pronouns

Lemma

Treebank	Wordform	Lemma	Lemma msd
Estonian: EWT	meie	mina	Pron.Pers.Sg1.Nom
Estonian: EDT	meie	mina	Pron.Pers.Sg1.Nom
North Saami: Giella	midjiide	mun	Pron.Pers.Sg1.Nom
Finnish: TDT	meillä	minä	Pron.Pers.Sg1.Nom
Finnish: PUD	meillä	**minä**	Pron.Pers.Sg1.Nom
Finnish: FTB	meillä	**me**	Pron.Pers.**Pl1**.Nom
Erzya: JR	минек	мон	Pron.Pers.**Pl1**.Nom
Karelian	hyö	hyö	Pron.Pers.**Pl3**.Nom
Komi: IKDP	миян	ми	Pron.Pers.**Pl1**.Nom
Komi: Lattice	миян	ми	Pron.Pers.**Pl1**.Nom
Hungarian: Szeged	nekünk	mi	Pron.Pers.**Pl1**.Nom

NumeralIssues=Yes

NumForm=Letter vs Digit

(attested in the Estonian treebanks but nowhere else)

Universal Quantifier ‘both’ = ‘all two’ PronType=Tot|PronType=Ind

est_ mõlemas mõlema DET Case=Ine|Number=Sing|PronType=Tot
hun_ mindkét mindkét DET Definite=Def|PronType=Ind
krl_ molompih molompi PRON Case=Ill|Number=Plur

Talbanken: bägge	bägge	DET	Definite=Def|Number=Plur|PronType=Tot
SynTagRus: обоим	оба	NUM	Case=Dat|Gender=Masc

Copula

- North Sámi, Estonian, Hungarian, Finnish and Karelian all have free copulas
  - Used differently, but regularly
- In Erzya copula can fuse into the stem with no clear boundary

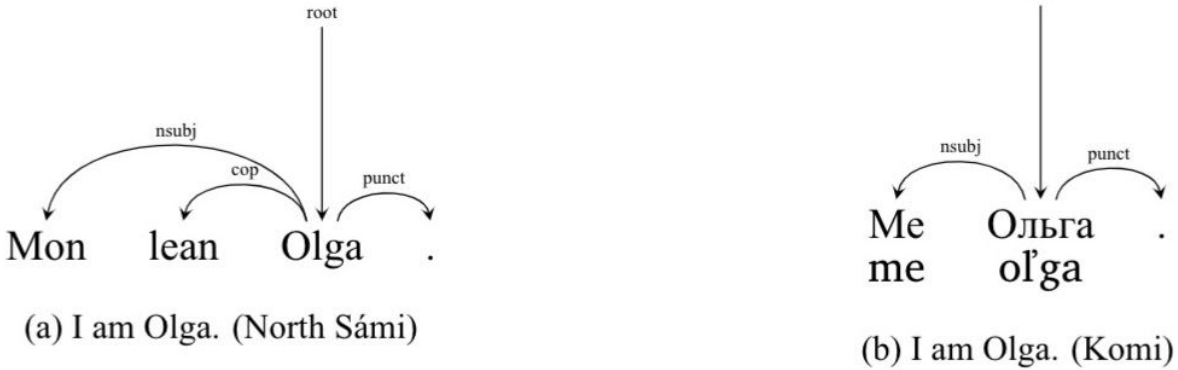


Figure 1: Example with and without copula

Third person singular may be seen as a ZERO formative

Personal pronoun tends to precede noun it is equated with

Locus of copula marking correlates to constituent stress.

(might be seen as contrastive stress)

(a) I'm OLGA. OR My name is OLGA. (Erzya) (b) I'm Olga. OR My name's Olga. (Erzya)

Figure 2: Distinguishing Erzya Subject

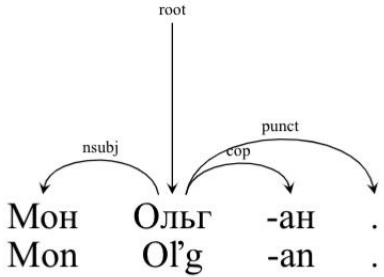


Figure 3: I'm OLGA. (Erzya)

Participles and features

- Deverbal nouns can be treated as nouns or verbs
- This decision has high impact to their dependencies too
- We compared parallel sentences previously discussed by Pirinen & Tyers (2016)

Example ‘I see the running man’

Language	Sentence	Features
North Saami	Oainnán **viehkki** dievddu.	Tense=Pres|VerbForm=Part
Erzya	Неян **чийниця** цёранть.	Case=Nom|Definite=Ind|Number=Sing Tense=Pres|VerbForm=Part
Finnish	Näen **juoksevan** miehen.	Case=Gen|Number=Sing|PartForm=Pres VerbForm=Part|Voice=Act
Estonian	Näen **jooksvat** meest.	Case=Par|Degree=Pos|Number=Sing Tense=Pres|VerbForm=Part|Voice=Act
Hungarian	Látom a **futó** embert.	‘ADJ’ _
Komi-Zyrian	Аддза **котралысь** мортöс.	PartForm=Pres|VerbForm=Part|Voice=Act

Example ‘I see the running man’

Language	Sentence	Agreed features?
North Saami	Oainnán **viehkki** dievddu.	**Tense=Pres|VerbForm=Part**
Erzya	Неян **чийниця** цёранть.	**Tense=Pres|VerbForm=Part**
Finnish	Näen **juoksevan** miehen.	**Tense=Pres|VerbForm=Part**
Estonian	Näen **jooksvat** meest.	**Tense=Pres|VerbForm=Part**
Hungarian	Látom a **futó** embert.	‘ADJ’ _
Komi-Zyrian	Аддза **котралысь** мортöс.	**Tense=Pres|VerbForm=Part**

Is there agreement up to this point? Can we document this agreement explicitly?

Other phenomena discussed in the paper

- Case names in different languages
- Use of indirect objects and obliques
- Use of feature Aspect in individual treebanks
- Number marking
- Marking of evidentiality

Conclusions

- Grammatical features specific to Uralic languages largely covered already
- Many language specific solutions originate from:
  - Traditional descriptions
  - Existing NLP tools (tagsets and conventions used)

- Even if everything were carefully checked against other treebanks, differences between them would make the task unclear
- With smaller treebanks harmonization-tasks still easily manageable
- One way or another, solution probably lies in **documentation**

Merci! Aitäh! Kiitos! Аттьö!
Köszönöm! Giitu! Tay!
Сюкпря! Thank you!