Négy hatás alatt álló nyelv

Korpuszépítés kis uráli nyelvekre

Simon Eszter

MTA Nyelvtudományi Intézet

Az előadás vázlata

1. Bevezetés
2. Szöveggyűjtés
3. Szövegfeldolgozás
4. A korpusz felépítése
5. Jövőbeli tervek

Bevezetés

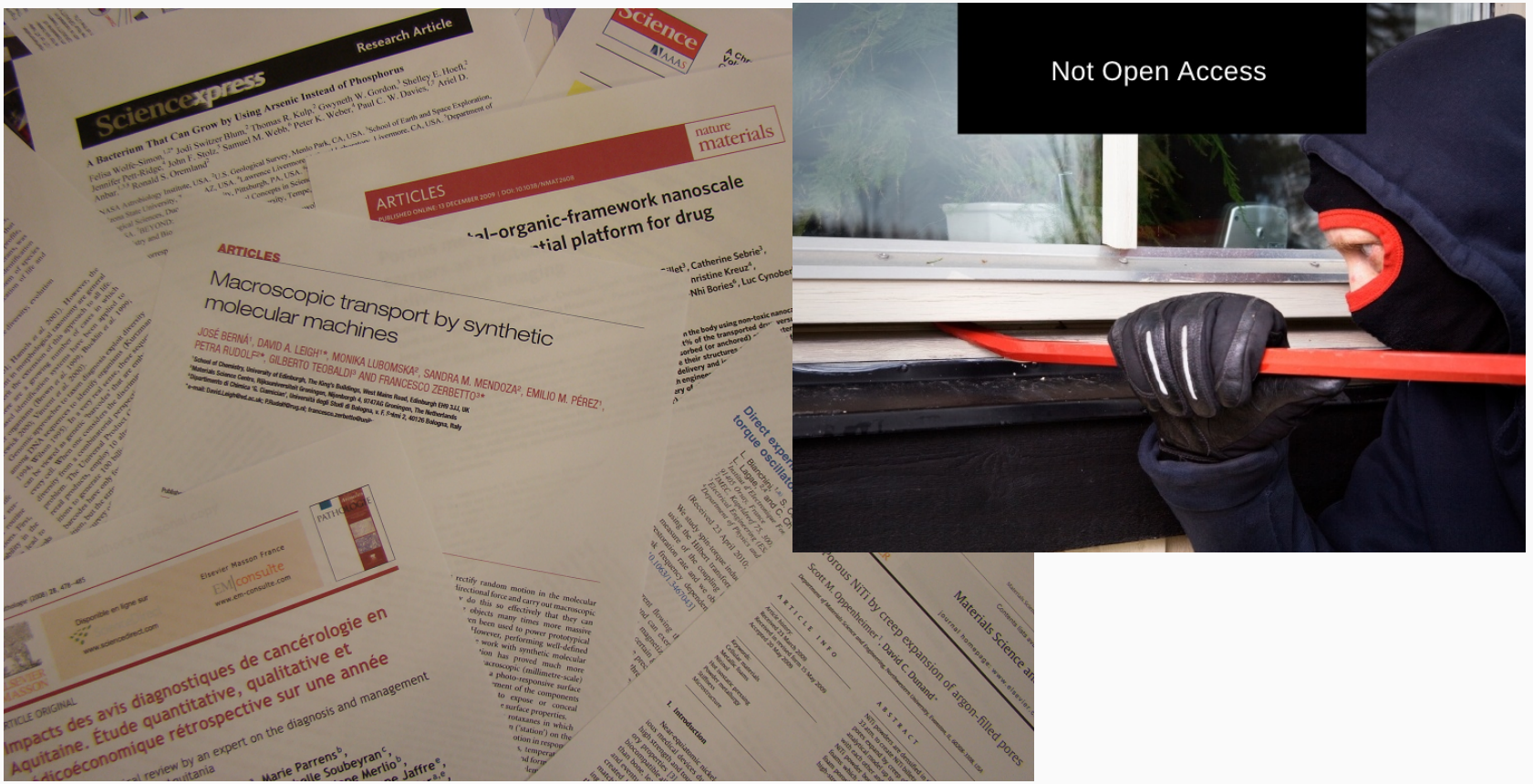
Uralisztikai kutatások 1.

a kutató terepmunkára megy valahova Oroszországba

Uralisztikai kutatások 2.

hazatér egy adag audió- és/vagy videófájllal, amit később feldolgoz a saját elképzeléseinek és céljainak megfelelően

Uralisztikai kutatások 3.



az adathalmazon kikutatott eredményeket publikálja, de az adathalmazt nem teszi publikusan hozzáférhetővé

Uralisztikai kutatások 4.

a kutató a beszélt nyelvi anyagot valami saját lejegyzési rendszer alapján jegyezte le, amit rajta kívül senki nem használ, és nem is ismer; dokumentáció, ami alapján meg lehetne fejteni a kódot, általában nincs, ha mégis van, akkor nincs publikálva, ha mégis, akkor nem angolul

különféle szövegszerkesztőkben, különféle házilag készített fontkészletekkel összeeszkábált, a strukturáltságnak látszatát kelteni sem igyekvő dokumentumok születnek

Célkitűzés

- egy nyelvi annotációt tartalmazó,
- sztenderd eszközökkel feldolgozott, és
- sztenderd formában,
- szabadon elérhető,
- strukturált adatbázis létrehozása

a számítógépes nyelvészet eszköztára jól használható erre a célra

A projekt

- *Az uráli nyelvek mondattanának változása aszimmetrikus kontaktushelyzetben*
- 2016. február – 2017. július
- MTA Nyelvtudományi Intézet
- interdiszciplináris csapat: kutatók a finnugor, a nyelvtechnológiai és az elméleti nyelvészeti osztályról
- projektvezető: É. Kiss Katalin
- NKFI-projekt (ERC_HU_15 118079)

Célok

A majdani ERC-projekt célja:

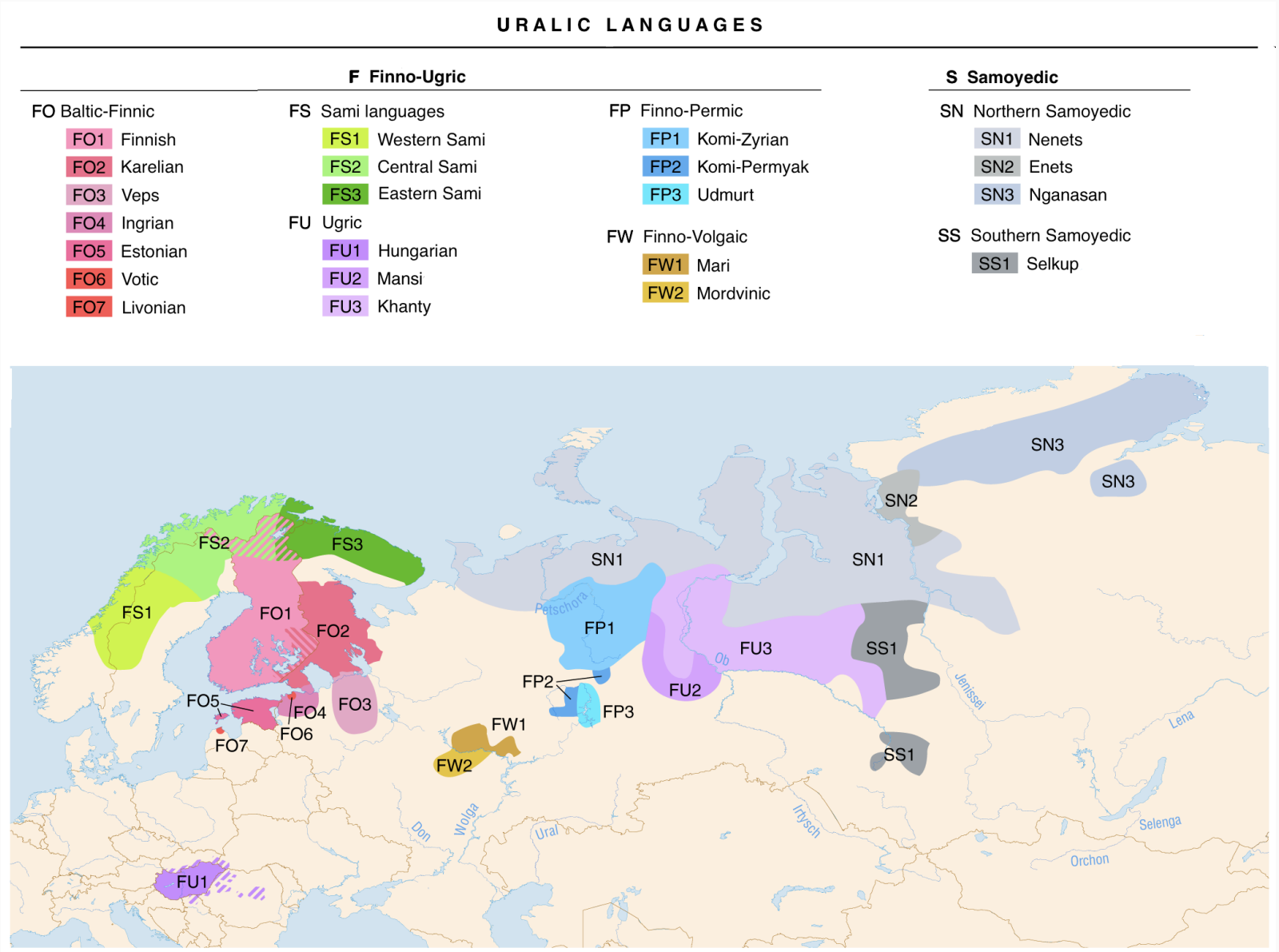
- elméleti: az orosz nyelv hatása négy oroszországi uráli nyelv szintaxisára (udmurt, tundrai nyenyec, szinjai és szurguti hanti)
- számítógépes: annotált korpusz létrehozása

A pilot projekt célja:

- a majdani ERC-projekt elméleti és módszertani alapjainak a kidolgozása
- egy pilot adatbázis építése:
  - 4000 token/kor/nyelv
  - IPA-átirat
  - teljes morfológiai elemzés
  - angol fordítás

Szöveggyűjtés

A vizsgált nyelvek



Az EGIDS (Expanded Graded Intergenerational Disruption Scale)

0	Nemzetközi	angol
1	Nemzeti	magyar
2	Regionális	
3	Kereskedelmi	
4	Oktatási	
5	**Írott**	udmurt
6a	Életerős	
6b	**Veszélyeztetett**	tundrai nyenyec, hantik
7	Nyelvcsere	
8a	Haldokló	
8b	Majdnem kihalt	
9	Alvó	
10	Kihalt	ógörög

A vizsgált nyelvek

Udmurt:

- írott (5) kategória: napi szinten használják, és létezik egy sztenderd irodalmi változata, de az nem annyira terjedt el
- Udmurtia egyik hivatalos nyelve
- vannak udmurt blogok, napi szinten keletkezik elektronikus udmurt szöveg
- udmurt Wikipédia (1000+ cikk)

Többiek:

- veszélyeztetett (6b) kategória: csak informális körben használják, alacsony presztízs
- nincs Wikipédia
- nincs napi sajtó, nem keletkezik elektronikus szöveg

A korpuszépítés kritériumai

- **elsődleges adatok:** olyan kommunikációs eseményekből származó nyelvi adatok, amelyek a hétköznapi nyelvhasználatot tükrözik
- **teljességre törekvés:**
  - minél több társadalmi osztályt, kort, nemet, műfajt és dialektust reprezentáljunk
  - metaadatok
  - az eredeti felvétel megőrzése, hogy a lejegyzések ellenőrizhetők legyenek
- **egységesség és összehasonlíthatóság:** nemzetközi sztenderdek
  - Unicode
  - IPA
  - Leipzig Glossing Rules
  - UTF-8 kódolású tsv fájlok

Források

	Régi	Új
Írott	folklór szövegek a 19. század végéről – 20. század elejéről	újabb gyűjtésű folklór szövegek & blogok, sajtó
Beszélt	$\emptyset$	terepmunka

Szövegfeldolgozás

Írás, transzkripció, transzliteráció

Írás:

- cirillalapú ábécék

Transzkripció (Finno-Ugric Transcription, FUT):

- szinjai hanti: Steinitz, RME
- szurguti hanti: Csepregi
- udmurt: Munkácsi, Wichmann
- tundrai nyenyec: Hajdú, Mus

Transzliteráció:

- IPA

Az eredeti szöveg előállítása

beszkennelt könyv → OCR → kézi javítás → eredeti szöveg

egységes karaktertábla: minden nyelv minden lejegyzési, átírási és írásrendszerének minden karaktere szerepel a Unicode-kódjával, -nevével és Prószéky-kódjával

- ezekkel a karakterekkel történik a hangzó szövegek lejegyzése,
- ezekre a karakterekre tanítjuk be az optikai karakterfelismerőt,
- ezekre a karakterekre normalizáljuk a különböző forrásokból származó szövegeket,
- és ezek szolgáltatják a különböző irányú konverziók bemeneti és kimeneti karakterállományát is

Karakterszintű normalizálás

különböző források $\rightarrow$ UTF-8 kódolású plain text fájlok

normalizálás:

- nem Unicode-karakterek lecserélése Unicode-karakterekre
- idegen nyelvű részek eltávolítása
- latin karakterek cirillre cserélése a cirill szövegben, pl. ван != ван

Konverzió

Összesen 12 konverziós irány:

szinjai hanti: Steinitz2IPA, Steinitz2RME
szurguti hanti: cirill2Csepregi, Csepregi2IPA
udmurt: Munkácsi2IPA, Wichmann2IPA, IPA2cirill, cirill2IPA
tundrai nyenyec: Hajdú2Mus, Hajdú2IPA, Hajdú2cirill, cirill2IPA

átírási szabályok → kiterjesztett reguláris kifejezéseket tartalmazó helyettesítési parancsok → `sed -f`

Szabályalapú rendszer:

Hátrány:	Előny:
nyelv- és irányfüggő	magas pontosság
könnyű hibázni	

Morfológiai elemzés

A cél: angol glosszázás az általunk létrehozott glosszázási rövidítéseket tartalmazó táblázat alapján

A táblázat az alábbi nemzetközi sztenderdek alapján készült:

- Leipzig Glossing Rules kódtábla
- a Wikipédia *Glossing abbreviations* című oldalának kódjai
- kurrens szakirodalom

Leképezés és konverzió:

az elérhető morfológiai elemzők kimenete → a mi kódtáblánk

Morfológiai elemzők

	MorphoLogic	Giellatekno
Udmurt	x	x
Tundrai nyenyec	-	x
Szinjai hanti	x	-
Szurguti hanti	-	-

Zipf törvénye alapján: a néhány leggyakoribb szó lefedi a teljes szöveg nagy százalékát → a min. ötször előforduló szavak kilistázása → elemzések hozzájuk rendelése kézzel → a szöveg több mint 60%-ához automatikusan hozzárendelődik az elemzés

A korpusz felépítése

Szövegfeldolgozottsági szintek

- eredeti szöveg
  - cirill
  - FUT lejegyzés(ek)
  - **IPA**
- **morfológiai információk**
  - lemma
  - szófajkód
  - inflexiós kódok
- fordítás
  - **angol**
  - orosz
  - német
  - magyar

A korpusz felépítése

YRK Hajdú:	jā	mīdaxana	amkerta	jaŋkūwi
YRK Mus:	ja	midaxana	amkerta	jaŋkuwi
YRK IPA:	ja	mi:daxana	ămkerta	jăŋkuwi
YRK cirill:	я	мыдахана	амкэрта	яңкувы
lemma:	я	мы	ңамгэ	яңгось
szófaj:	N	Ptcp	Pron.neg	V
glossza:	earth	create.IPFV.PTCP.LOC	what.CONC	neg.EX.INFER

ENG:	when the earth was created, there was nothing
GER:	zur zeit der erschaffung der erde gab es nichts
HUN:	a föld teremtésének idején nem volt semmi

Jövőbeli tervek

Jövőbeli tervek

- a projekt folytatása az ERC keretein belül
- a korpusz bővítése
- a projekt során előálló szöveges és feldolgozó erőforrásokat szabadon elérhetővé tesszük
  - a projekt weboldalán: `http://www.nytud.hu/oszt/elmnyelv/urali`
  - online lekérdező felületen keresztül
  - egy nemzetközi nyelvi archívumon keresztül: Documentation of Endangered Languages (DOBES) by The Language Archive
- a konvertereket átírjuk → a Giellatekno részévé szeretnénk tenni
- a Giellatekno tundrai nyenyec elemzőjének fejlesztése: új, nyelvjárási szótári elemek & a nyelvtanfájlok ápdételése a legújabb nyelvtan alapján

Köszönöm a figyelmet!

simon.eszter@nytud.mta.hu

http://www.nytud.hu/oszt/elmnyelv/urali