Lexikai erőforrások automatikus előállítása kisebbségi finnugor nyelvekre

Simon Eszter, Mittelholcz Iván, Ferenczi Zsanett

XIV. MSZNY, Szeged, 2018. január 19.

MTA Nyelvtudományi Intézet

Az előadás vázlata

1. Bevezetés
2. Automatikus szótárépítés
3. Wiktionary-szócikkek generálása
4. Összegzés

Bevezetés

A projekt

Finnugor nyelvű közösségek nyelvtechnológiai támogatása online tartalmak létrehozásában

- OTKA No. FNN 107885
- projektvezető: Váradi Tamás
- 2013. szeptember – 2018. február
- partnerek:
  - MTA Nyelvtudományi Intézet
  - Helsinki Egyetem

Digitális revitalizáció I.

Mi van?

- változások: az információs igény nagy része online forrásokból van kielégítve & a kommunikációs technológia a személyes életünkben is fontos szerepet játszik → a nyelv a vivőanyag → fontos, hogy gyorsan adaptálódjon az új helyzetekhez
- a digitálisan életképes nyelveknek nagy előnye van ⇔ a kisebbségi nyelveket beszélő közösségek a legérzékenyebbek a változásokra

Mi mit tudunk tenni?

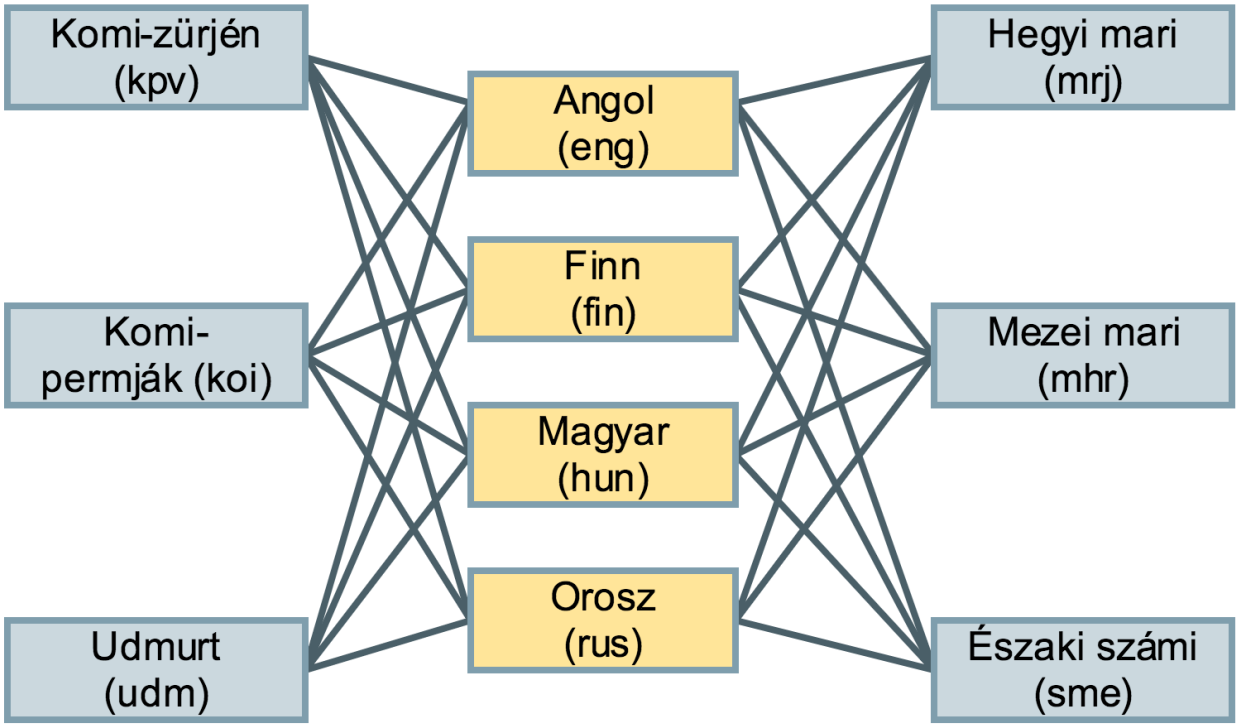
- kétnyelvű szótárakat állítunk elő → újabb szópárokkal gazdagítjuk az interneten fellelhető fordítási párok számát
- Wiktionary-szócikkeket generálunk → feltöltjük őket

Digitális revitalizáció II.

Mi ennek az értelme?

- a szótári elemek a Wiktionary különböző nyelvű változataiban összekapcsolhatók & az interwiki linkek pedig a Wikipédia felé biztosítják az átjárást → a nyelvközösségek gazdag lexikai anyaghoz férnek hozzá
- ezzel támogatjuk a veszélyeztetett finnugor nyelvű közösségeket a digitális revitalizációban & nyelvi erőforrásokkal járulunk hozzá a további fejlesztésekhez
- hozzájárulunk a nyelvi sokszínűség fenntartásához

A nyelvpárok



EGIDS (Expanded Graded Intergenerational Disruption Scale)

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

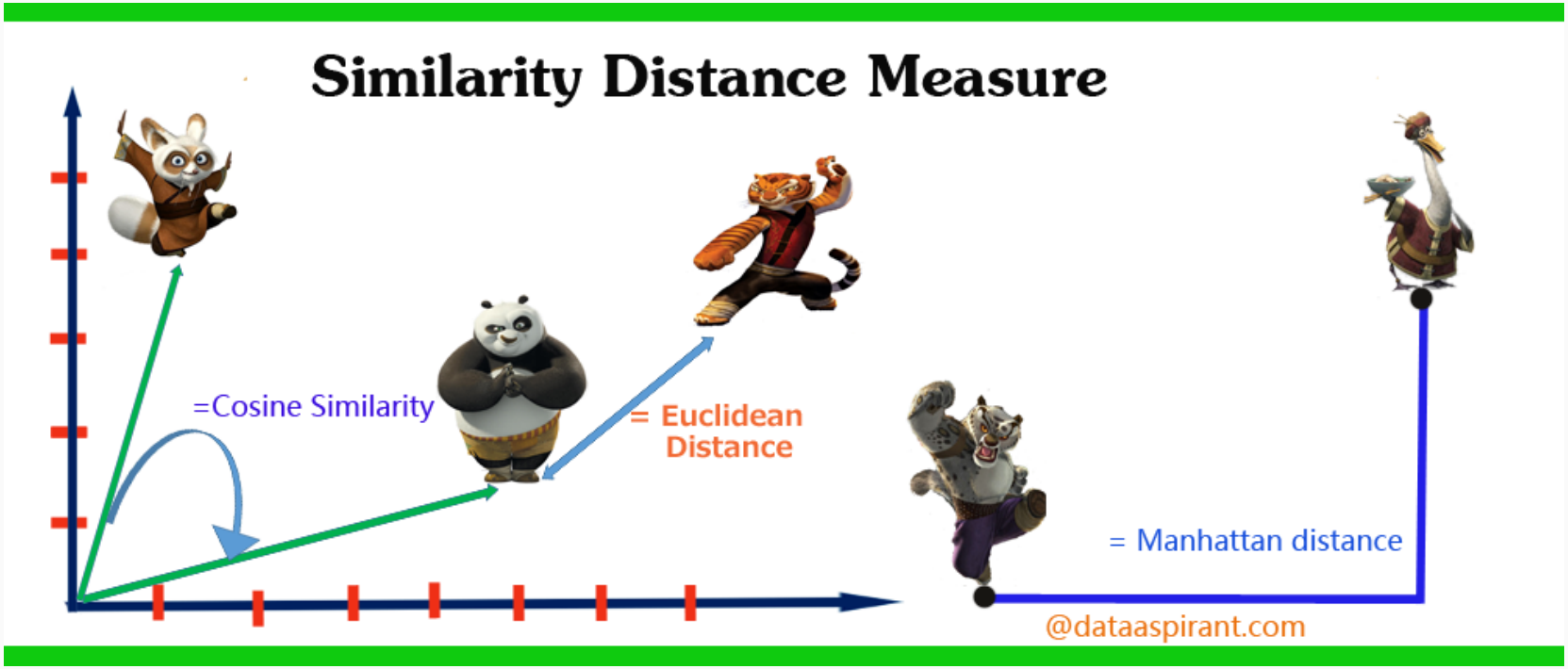
A vizsgált nyelvek

nyelv	ISO	EGIDS	népesség	terület	írás
északi számi	sme	2	20.000	Norvégia, Svédország, Finnország	latin
mezei mari	mhr	4	470.000	Oroszország	cirill
hegyi mari	mrj	5	30.000	Oroszország	cirill
komi-zürjén	kpv	5	156.000	Oroszország	cirill
komi-permják	koi	5	63.000	Oroszország	cirill
udmurt	udm	5	340.000	Oroszország	cirill

Automatikus szótárépítés

Sztenderd szótárépítési módszerek

- párhuzamos vagy összevethető korpuszokból
- a forrás- és célnyelvi szavakat reprezentáló vektorok
- hasonlóságszámítás



hátrányuk:

nagy mennyiségű előfeldolgozott szöveget igényelnek

Alternatív módszerek

közösség által épített nyelvi erőforrások használata: Wikipédia, Wiktionary

Wiktionary

- többnyelvű szótár
- a Wikipédia testvérprojektje
- szabadon elérhető (CC-BY-SA 3.0 és GNU Free Documentation License)
- minden nyelv minden szavát tartalmazni akarja
- sok nyelven elérhető, és a szócikkek mindig az adott Wiktionary nyelvén szerepelnek

Wiktionary

Hungarian [edit]

Pronunciation [edit]

- IPA(key): ['hoː]
- Audio 0:00 MENU

Etymology 1 [edit]

From Proto-Uralic **kume* ("thin snow"). Cognates include Tundra Nenets хав" (*χāβ*, "a thin crust of snow").

Noun [edit]

hó (*plural* **havak**)

1. snow

Declension [edit]

Inflection (stem in *-a-*, back harmony) [more ▼]
Possessive forms of *hó* [more ▼]

Wikipédia-címszópárok

WIKIPEDIA
The Free Encyclopedia

Languages

Afrikaans
Alemannisch
العربية
Bân-lâm-gú
Беларуская
Беларуская (тарашкевіца)
Български
Brezhoneg
Català
Čeština
Dansk
Deutsch

Szeged

From Wikipedia, the free encyclopedia

Coordinates: 46.255°N 20.145°E

Not to be confused with Seget.

Szeged (Hungarian pronunciation: [ˈsɛgɛd] (listen); see also other alternative names) is the third largest city of Hungary, the largest city and regional centre of the Southern Great Plain and the county seat of Csongrád county. The University of Szeged is one of the most distinguished universities in Hungary.

The famous Szeged Open Air (Theatre) Festival (first held in 1931) is one of the main attractions, held every summer and celebrated as the Day of the City on May 21.

Szeged

Urban county

az interwiki linkek segítségével kétnyelvű szótárakat hoztunk létre

Wiktionary-alapú módszerek I.

wikt2dict

`https://github.com/juditacs/wikt2dict`

Parszolás

- parszoltuk az angol, magyar, finn és orosz Wiktionary-t
- a fordítási táblákból kinyertük a fordítási megfelelőket

Translations [edit]

±association of tradespeople [hide ▲]

Select targeted languages

- Armenian: գիլդիա (hy) (gildia), համքարութիւն (hy) (hamkʿarutʿyun)
- Bulgarian: еснаф (bg) *m* (esnaf), гилдия (bg) *f* (gildija)
- Chinese:
  - Mandarin: 行會 (zh), 行会 (zh) (hánghuì), 同業公會 (zh), 同业公会 (zh) (tóngyè gōnghuì), 公會 (zh), 公会 (zh) (gōnghuì)
- Czech: cech (cs) *m*
- Danish: lav (da)
- Dutch: gilde (nl) *n*
- Esperanto: gildo (eo)
- Finnish: kilta (fi)
- French: guilde (fr)
- German: Gilde (de) *f*, Zunft (de) *f*
- Hungarian: céh (hu)
- Ido: gildo (io)
- Irish: gild *m*
- Italian: gilda, fraglia *f*
- Japanese: ギルド (girudo)
- Lithuanian: gildija
- Norwegian: laug
- Polish: gildia (pl), cech (pl)
- Portuguese: guilda (pt) *f*
- Russian: ги́льдия (ru) *f* (gílʹdija)
- Serbo-Croatian: ceh (sh) *m*
- Slovene: ceh (sl)
- Spanish: germanía (es) *f*, gremio (es) *m*
- Swedish: skrå (sv) *n*, gille (sv) *n*
- Tagalog: gremyo
- Welsh: urdd *f*

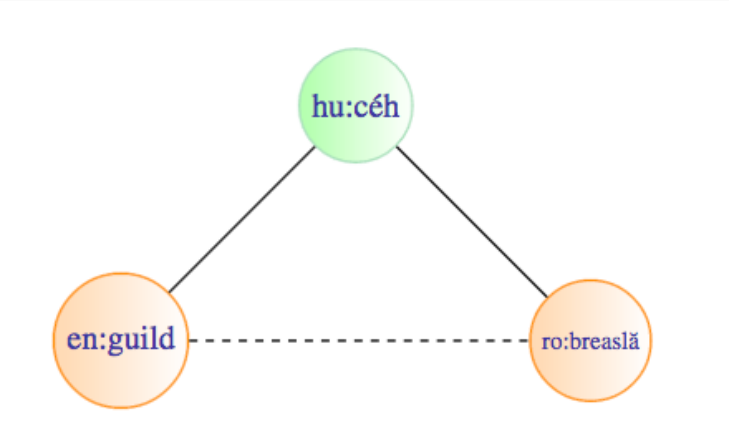
Add translation : Preview translation More

Script code: (e.g. Cyrl for Cyrillic, Latn for Latin)

Wiktionary-alapú módszerek II.

Háromszögelés

- új kapcsolatokat hoz létre már meglevő fordítási párokból
- két elem vsz. egymás fordítása, ha mindkettő egy harmadiknak a fordítása



Kiértékelés

módszer	össz (#)	ok-ok (%)
W2D ext	16.854	71,76
W2D tri	1.965	56,61
WikiTitle	23.066	54,11
KDE4	8.401	27,57

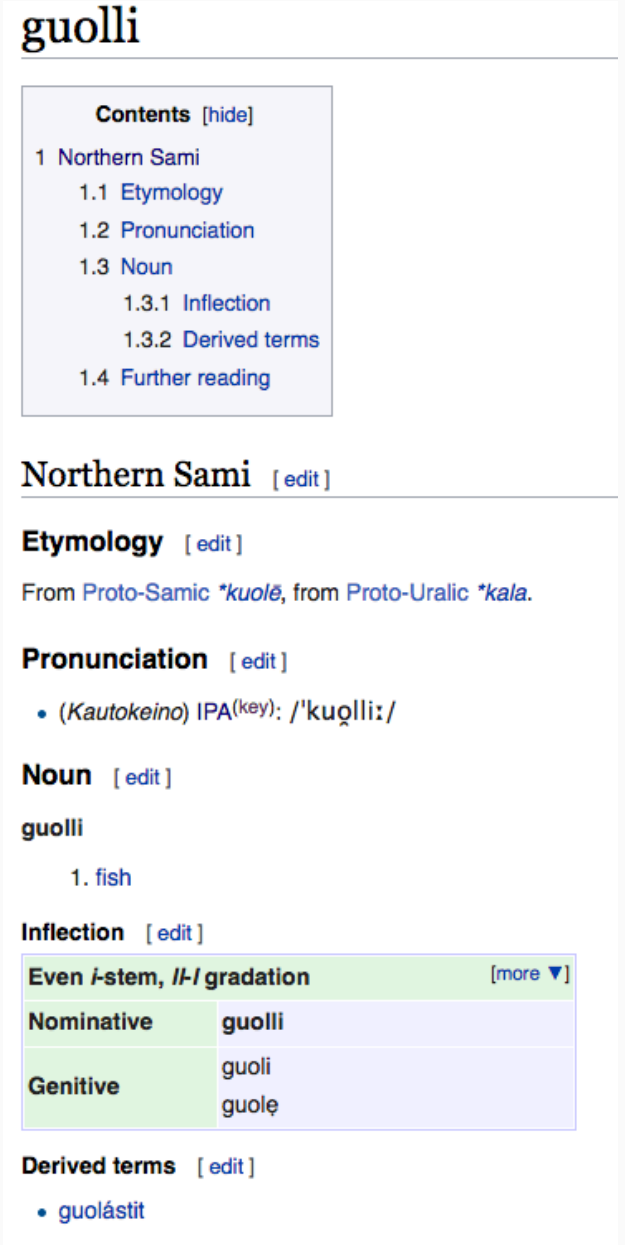
Az egyes módszerek összehasonlítása.

KDE4: az Opus korpuszból letöltött szótárak, sztenderd szótárépítési módszerrel készültek, az sme–{eng,fin,hun} nyelvpárokra

Wiktionary-szócikkek generálása

Wiktionary-szócikkek létrehozása

```
==Northern Sami==
===Etymology===
From {{inh|se|smi-pro|*kuolē}},
from {{inh|se|urj-pro|*kala}}.
===Pronunciation===
* {{se-IPA}}
===Noun===
{{se-noun}}
# [[fish]]
====Inflection====
{{se-infl-noun-even|guolli}}
====Derived terms====
* {{l|se|guolástit}}
[[Category:se:Fish]]
```



guolli

Contents [hide]

1 Northern Sami
- 1.1 Etymology
- 1.2 Pronunciation
- 1.3 Noun
  - 1.3.1 Inflection
  - 1.3.2 Derived terms
- 1.4 Further reading

Northern Sami [edit]

Etymology [edit]

From Proto-Samic **kuolē*, from Proto-Uralic **kala*.

Pronunciation [edit]

- (*Kautokeino*) IPA(key): /ˈkuo̯lliː/

Noun [edit]

guolli

1. fish

Inflection [edit]

Even *i*-stem, *II-I* gradation		[more ▼]
Nominative	guolli	
Genitive	guoli guolę	

Derived terms [edit]

- guolástit

A szófaji kategória meghatározása

morfológiai elemzők:

- koi, kpv, mhr, mrj, sme, udm; fin, rus: Giellatekno
- hun: `emMorph`
- eng: `hunmorph`

egyértelműsítés:

1. morfológiai információk alapján történő szűrés
2. horizontális összehasonlítás
3. vertikális összehasonlítás

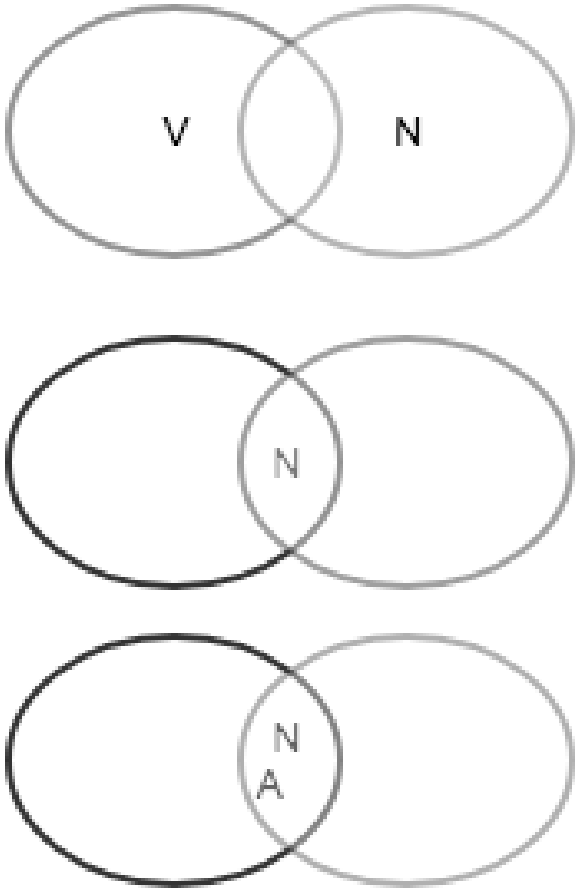
Morfológiai információk alapján történő szűrés

a validált protoszótárak csak szótári alakokat tartalmaznak

- szó = lemma
- névszók: Sg Nom, igék: Sg3 Pres Indef Indicative
- többszavas kifejezések: utolsó tag
- szűkebb kategória: Prop N $\rightarrow$ Prop

Horizontális és vertikális összehasonlítás

Horizontális összehasonlítás



Vertikális összehasonlítás

S	POS	T	POS
ань	?	female	A, **N**
ань	?	mother	**N**, V
ань	?	woman	**N**, V

Fonetikai átírás

- koi, kpv, mhr, mrj, udm: a Mari Web Project automatikus átírási eszköze (`http://www.univie.ac.at/maridict/site-2014/transcription-general.php?int=0`)
- sme: Giellatekno `text2ipa` → FST (→ később töröltük)

nincs átírás:

- tulajdonnevekhez
- számot tartalmazó szavakhoz
- északi számi szavakhoz

A szócikkek előállítása és feltöltése

- a legfrissebb Wiktionary-dumpokban ellenőrizzük, hogy az adott szó létezik-e már → ha igen, akkor nem generálunk hozzá szócikket
- egy forrásnyelvi szóhoz több célnyelvi szó tartozik, és ezek szófaja ugyanaz → egy fejléc alá kerülnek
- egy forrásnyelvi szóhoz több célnyelvi szó tartozik, de nem ugyanaz a szófajuk → külön fejlécek alá kerülnek
- egy forrásnyelvi szó több nyelven ugyanaz → egyesítjük a szócikkeket
- minden elemet összerakva teljesen automatikusan generáljuk a szócikkeket
- `Pywikibot --safe`
- a botok használata korlátozva van, engedélyt kell kérni

Kiértékelés

nyelv	hasznos (#)	maradék (#)	wikt (#)	új (#)	növ (%)
kom	1.025	699	152	225	148,02
chm	2.589	1.634	34	869	2555,88
sme	4.509	2.392	206	2.000	970,87
udm	1.198	739	128	420	328,13

A létrehozott szócikkek kiértékelése az egyes nyelvekre a Wikiszótárban.

Összegzés

Tanulságok és tervek

- a sztenderd szótárépítési módszerek a kevés erőforrással rendelkező nyelvekre nem annyira hatékonyak → a közösségileg épített erőforrások jól használhatók
- a Wikiszótárban található FU nyelvű szavak számát megsokszoroztuk
- a Wikisanakirjába való feltöltés folyamatban van
- a Викисловарь-ba és a Wiktionary-be való feltöltéshez még be kell szerezni az engedélyeket
- a szótáraink a Giellatekno erőforrásaiba is be fognak kerülni
- tervezzük a létrehozott lexikai erőforrások RDF-esítését → a Linguistic Linked Open Data részévé akarjuk tenni
- egy honlapon közzétesszük a létrehozott erőforrásokat

Köszönjük a figyelmet!

simon.eszter@nytud.mta.hu
mittelholcz.ivan@nytud.mta.hu
ferenczi.zsanett@nytud.mta.hu