Szócikképítés a Wiktionaryben lépésről lépésre

Ferenczi Zsanett

2018. február 13.

MTA Nyelvtudományi Intézet

Az előadás vázlata

1. Bevezetés
2. Wiktionary
3. Szócikkgenerálás lépései
4. Összegzés

Bevezetés

Bevezetés

- protoszótárakból Wiktionary szócikkek
- többletinformációkkal való kiegészítés
- automatikus előállítás
- automatikus feltöltés

Wiktionary

Wiktionary

- sok nyelven elérhető
- a szócikkek mindig az adott Wiktionary nyelvén szerepelnek
- feltöltés a Wiktionary négy nagyobb nyelvű kiadásába (eng, fin, hun, rus) $\rightarrow$ célnyelvek

Wiktionary szócikk

```
==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

Wikitext

- Fejlécek → `=`
- Sablonok → `{{}}`
- Felsorolás, lista → `#, *`
- Linkek → `[[]]`

Szócikkgenerálás lépései

A protoszótárak

Kétnyelvű szótárak

- forrásnyelvi szó (koi, kpv, mhr, mrj, sme, udm) szótári alakban
- célnyelvi szó (eng, fin, hun, rus) szótári alakban

sme	fin
amas	outo
amerihkálaš	amerikkalainen
analiisa	analyysi
analiisa	tarkastelu
analiisa	tutkimus
anatomiija	anatomia
animašuvnna	animaatio

A szócikk felépítése

Kötelező információk:

- forrásnyelv
- forrásnyelvi szó
- célnyelvi szó
- szófaji kategória

Egyéb lehetséges információk:

- fonetikai átírás
- etimológia
- szinonimák, antonimák
- ragozási táblák
- stb.

A szófaji kategória meghatározása

morfológiai elemzők:

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

egyértelműsítés:

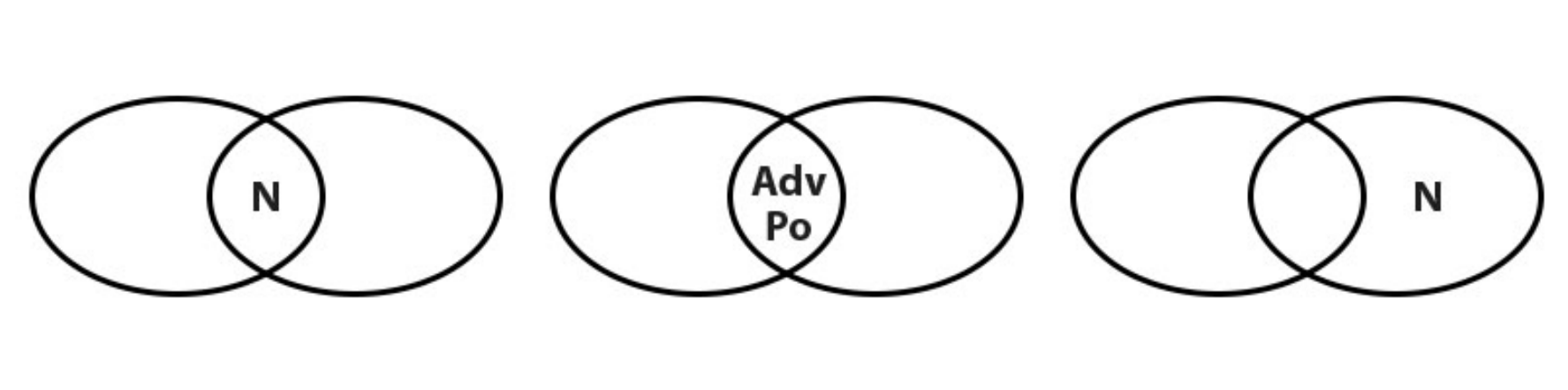
1. morfológiai információk alapján történő szűrés
2. horizontális összehasonlítás
3. vertikális összehasonlítás

Morfológiai információk alapján történő szűrés

a validált protoszótárak csak szótári alakokat tartalmaznak

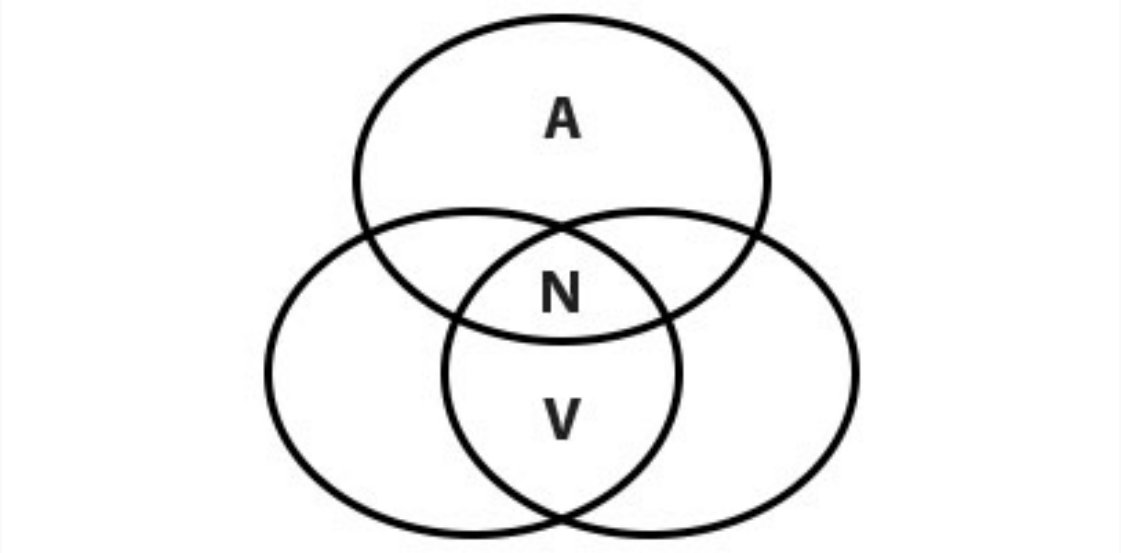
- szó = lemma
- névszók: Sg Nom, igék: Inf
  biebmu
  biebmat +V+TV+Imprt+Du1 **biebmu** +N+Sg+Nom
- többszavas kifejezések: utolsó tag
  (pl. nubbi **máilbmesoahti** = második **világháború**)
- szűkebb kategória: Prop, N $\rightarrow$ Prop

Horizontális összehasonlítás



S	POS	T	POS
loahppa	**N**	loppu	Adv, A, **N**
gasku	**Po, Adv**, Pr	keskellä	**Po, Adv**, Pcle
duođáštus	?	todistus	**N**

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` ($\rightarrow$ később töröltük)

nincs átírás:

- tulajdonnevekhez
- számot tartalmazó szavakhoz
- északi számi szavakhoz

Generálás előtt

S	T	POS	IPA
луд	kenttä	N	lud
луд	pelto	N	lud
юдэс	kappale	N	judes
уром	ystävä	N	urom

- a legfrissebb Wiktionary-dumpokban ellenőrizzük, hogy az adott szó létezik-e már → ha igen, akkor nem generálunk hozzá szócikket

A szócikkek előállítása

- ha a szófaj ugyanaz → egy fejléc alá
- ha a szófaj különböző → külön fejlécek alá

Unkari

Substantiivi

dob

1. rumpu

Verbi

dob

1. heittää, viskata

- egy forrásnyelvi szó több nyelven ugyanaz → egyesítjük a szócikkeket (pl. *köин* - komi-permják, komi-zürjén)

A szócikkek feltöltése

- minden elemet összerakva, teljesen automatikusan generáljuk a szócikkeket
- `Pywikibot --safe`
- a botok használata korlátozva van, engedélyt kell kérni
- Wikiszótárba és Wikisanakirjába való feltöltés már kész

Kiértékelés – Wikiszótár

nyelv	össz (#)	wikt (#)	közös (#)	új (#)	növ (%)
kom	699	152	35	664	436,84
chm	1.633	34	12	1.621	4.767,65
sme	2.392	206	146	2.246	1.090,29
udm	729	128	69	660	515,62
össz		520		5.191	998,27

A létrehozott szócikkek kiértékelése az egyes nyelvekre a Wikiszótárban.

Kiértékelés – Wikisanakirja

nyelv	össz (#)	wikt (#)	közös (#)	új (#)	növ (%)
kom	687	42	27	660	1.571,43
chm	1.903	443	213	1.690	381,49
sme	2.862	817	422	2.440	298,65
udm	828	55	45	783	1.423,64
össz		1.357		5.573	410,68

A létrehozott szócikkek kiértékelése az egyes nyelvekre a Wikisanakirjában.

Összegzés

Tanulságok és tervek

- a Wikiszótárban és a Wikisanakirjában található FU nyelvű szavak számát megsokszoroztuk
- 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 $\rightarrow$ a Linguistic Linked Open Data részévé akarjuk tenni
- egy honlapon közzétesszük a létrehozott erőforrásokat

Köszönöm a figyelmet!

ferenczi.zsanett@nytud.mta.hu