Automatikus szótárépítési módszerek finnugor nyelvekre

Ferenczi Zsanett

2020. június 25.

PPKE BTK Nyelvtudományi Doktori Iskola
Nyelvtechnológia Műhely
Témavezető: Simon Eszter

Vázlat

1. A kutatás bemutatása
2. Lexikográfiai adatbázis
3. Szócikkek felépítése
4. További feladatok

A kutatás bemutatása

A kutatás bemutatása

- *”Automatikus szótárépítési módszerek finnugor nyelvekre”*
- szótárépítési módszerek összehasonlítása
- finnugor nyelvek többnyelvű online szótára
- nyelvtanuló alkalmazás (CALL)
- létező nyelvtechnológiai eszközök közötti kompatibilitás biztosítása, új eszközök létrehozása

Eredmények

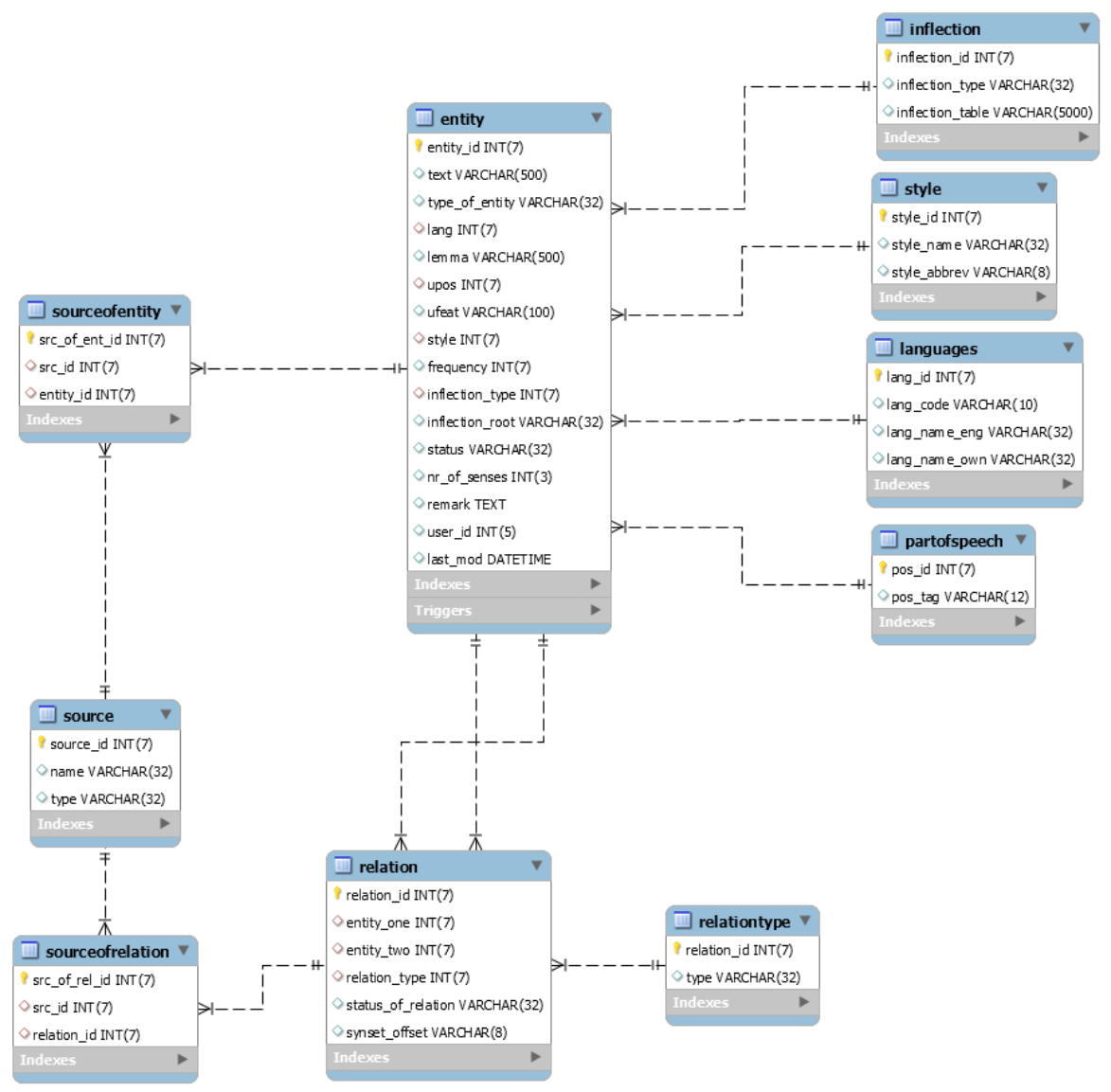
- **új eszközök**:
  - finn és magyar Wiktionaryk feldolgozása
  - finn és magyar WordNetek feldolgozása, összekötése
  - tremjugani és vaszjugani hanti > betűrendbe sorolás
- **szótár**:
  - szópárok kinyerése különböző módszerekkel
  - nyelvfüggetlen lexikográfiai adatbázis felépítése
  - lemmák morfológiai elemzése, szófaji taggelés
  - definíciók, inflexiós tőtípusok, példamondatok kinyerése

Lexikográfiai adatbázis

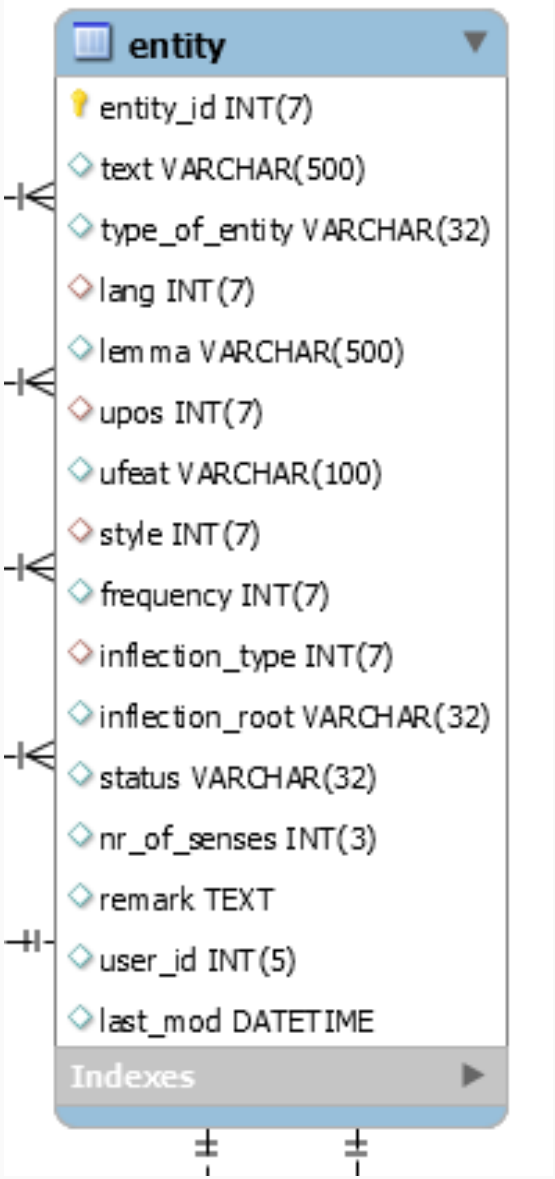
Lexikográfiai adatbázis

- cél: **többnyelvű szótár** adatainak elraktározása, összegyűjtött információk **redundanciamentes** tárolása
- minden nyelvi információt azonos módon kezelünk → **entitás**
- az entitások típusai: *lemma*, *szóalak*, *többszavas kifejezés*, *definíció*, *mondat*, stb.
- a jelentést az egyes entitások közötti relációkkal tudjuk megragadni
- a relációk közt megtalálhatók szemantikai relációk (mint *szinonima*, *antonima*), illetve különböző lexikográfiai relációk (pl. *fordítása*, *definíciója*, *példamondata*, stb.)
- az entitáspárok és entitások forrását eltároljuk

Az adatbázis



Az adatbázis - entitás



Szócikkek felépítése

Példa 1.

apu = noun

isä, iskä

(Apa megszólítása gyermeke részéről)

Apu szereti anyut.

Isä rakastaa äitiä.

szinonimák: apuci, apuka

apu + noun

segítség

Apu tuli viime hetkellä.

A segítség az utolsó pillanatban érkezett.

Példa 2.

ég noun

taivas

etimológia: *Ősi finnugor örökségünk: zürjén szünöd (‘levegő’), finn sää (‘időjárás’).*

szinonimák: égbolt, menny
összetételek: éghajlat, égtáj

ég verb

palaa

(Hő és fény kibocsátása közben pusztul, fogy.)

ragozás:

E/1	égek
E/2	égsz
E/3	ég
T/1	égünk
T/2	égtek
T/3	égnek
	…

szinonimák: lángol, parázslik

Példa 3.

daru noun

(madár) **kurki**

(szürke, gólyaszerű gázlómadár)

etimológia: *ogul tárov, osztják táreh, zürjén, votják turi ('daru').*

ragozás: ... darvak ...

(emelőgép) **nosturi**

ragozás: ... daruk ...

Összegyűjtött szópárok

finn és magyar szópárok

- Wiktionary - **Wikt2dict**
  - Ács et al. (2013) Wiktionaryt feldolgozó eszköze
  - általában lemmák, előfordulnak nem alapalakok is
  - két módon: *`extract`*, *`triangulate`*
  - finn, magyar és angol Wiktionary verziók
  - összesen **16 685 egyedi szópár**
- finn és magyar Wiktionary
  - saját fejlesztésű, Wiktionary dumpot feldolgozó eszköz
  - szópárokat, szófaji címkéket és tőtípust nyer ki
  - összesen **8 762 egyedi szópár**

Összegyűjtött szópárok

- WordNet
  - FinnWordNet (Lindén és Carlson (2010))
  - HuWN (Miháltz et al. (2008))
  - **99 780 szópár** (szófaji címkével)
  - 25 662 synsetpár

poikanen, kissanpentu **cica, kiscica, kölyökmacska, kismacska**

- OpusCorpus
  - párhuzamos korpusz
  - szavak nincsenek lemmatizálva
  - **832 696 szópár**

kissanpennuilla **macskakölyöknek**
kissanpennut **kiscicák**
kissanpentu **kiscica**

Szófaji címke

- **Wikt2dict** eszköz és **OpusCorpus** nem ad meg szófaji címkéket
- morfológiai elemzőket használva kell szófajt kinyerni
- egységes UD kimenetettel
- **omorfi** - finn szavakhoz
- **emMorph** - magyar szavakhoz, **emMorph2UD** eszköz
- minden elemzést megtartva

Definíciók

- magyar Wiktionaryről magyar szavakhoz
  - **27 154** egyedi szóhoz 31 469 definíció

pad Fából vagy kőből készült hosszú ülőke.

- finn Wiktionaryről finn szavakhoz
  - **102 638** egyedi szóhoz 127 786 definíció

- magyar WordNetről
  - **35 011** egyedi szóhoz 42 356 definíció

dolog Független, önmagában álló entitás.
dolog Valamilyen cselekedet, tett, magatartás.
dolog Elvégzendő feladat, munka.

Igei és névszói paradigmák és példamondatok

- példamondatok a WordNetről:
  - magyar: **22 562** synsethez
- paradigmák
  - finn Wiktionaryről: **24 121** finn szóhoz
  - magyar Wiktionaryről: **4 704** magyar szóhoz
  - tőtípusok sablonjai:

    poliszémia szócikk esetén:
    {{hu-fn-k1|poliszémi|á}}{{hu-birt-tok|poliszémiá}}

Inflexiós táblák

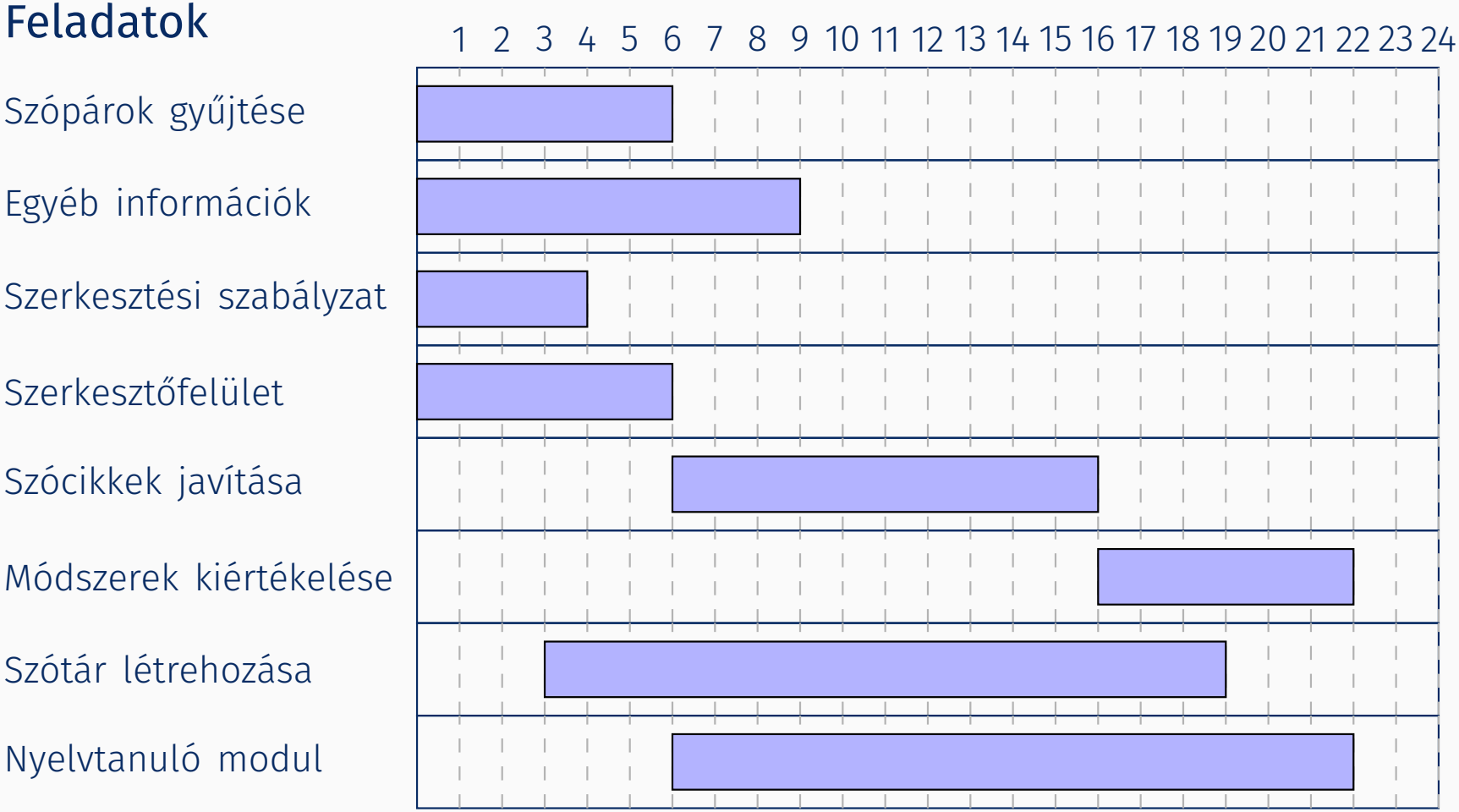
- a tőtípusokhoz tartozó inflexiós táblák kinyerése

hu-fn-k1 ragozása [becsuk]		
eset/szám	**egyes szám**	**többes szám**
alanyeset	hu-fn-k1	{{{1}}}{{{2}}}k
tárgyeset	{{{1}}}{{{2}}}t	{{{1}}}{{{2}}}kat
részes eset	{{{1}}}{{{2}}}nak	{{{1}}}{{{2}}}knak
-val/-vel	{{{1}}}{{{2}}}val	{{{1}}}{{{2}}}kkal
-ért	{{{1}}}{{{2}}}ért	{{{1}}}{{{2}}}kért
-vá/-vé	{{{1}}}{{{2}}}vá	{{{1}}}{{{2}}}kká
-ig	{{{1}}}{{{2}}}ig	{{{1}}}{{{2}}}kig
-ként	hu-fn-k1ként	{{{1}}}{{{2}}}kként
-ul/-ül	-	-
-ban/-ben	{{{1}}}{{{2}}}ban	{{{1}}}{{{2}}}kban
-on/-en/-ön	{{{1}}}{{{2}}}n	{{{1}}}{{{2}}}kon
-nál/-nél	{{{1}}}{{{2}}}nál	{{{1}}}{{{2}}}knál
-ba/-be	{{{1}}}{{{2}}}ba	{{{1}}}{{{2}}}kba
-ra/-re	{{{1}}}{{{2}}}ra	{{{1}}}{{{2}}}kra
-hoz/-hez/-höz	{{{1}}}{{{2}}}hoz	{{{1}}}{{{2}}}khoz
-ból/-ből	{{{1}}}{{{2}}}ból	{{{1}}}{{{2}}}kból
-ról/-ről	{{{1}}}{{{2}}}ról	{{{1}}}{{{2}}}król
-tól/-től	{{{1}}}{{{2}}}tól	{{{1}}}{{{2}}}któl

További feladatok

Kutatással kapcsolatos teendők

Feladatok



További feladatok - CALL

- feleletválasztós kérdések
  - pl. _____ *on iso kissa.*
    a) eläin **b)** hiiri **c)** koira **d)** tiikeri
  - definíciókból, példamondatokból
  - a célszóhoz hasonló választási lehetőségekkel

- mondatba illő szóalak megadása
  - pl. *Mitä nähtävyyksiä* _____ *(kaupunki) on?*
  - definíciókból, példamondatokból

- szókincsfejlesztés
  - forrás- és célnyelvi szavak összepárosítása
  - szinonimák keresése

Kutatási és oktatási modul

- ☒ Recenzió magyar nyelvű könyvről
- ☒ Magyar nyelvű lektorált kiadvány szerkesztése
- ☒ Tudománynépszerűsítő közlemény
- ☒ Egyéb elismert tudományos vagy tudománynépszerűsítő tevékenység
- ☒ Oktatási modul
- ☐ Saját kutatáson alapuló tudományos publikációk
- ☐ Konferenciaelőadások (AlkNyelvDok, MSZNY, IWCLUL, NLP4CALL)
- ☐ Poszterek

Bibliográfia

Judit Ács, Katalin Pajkossy, and Andras Kornai. Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pp. 52–58, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL *http://www.aclweb.org/anthology/W13-2507*.

Krister Lindén and Lauri Carlson. Finnwordnet–finnish wordnet by translation. *LexicoNordica–Nordic Journal of Lexicography*, 17: 119–140, 2010.

Márton Miháltz, Csaba Hatvani, Judit Kuti, György Szarvas, János Csirik, Gábor Prószéky, and Tamás Váradi. Methods and results of the hungarian wordnet project. In *Proceedings of The Fourth Global WordNet Conference*, pp. 311–321, 2008.

Köszönöm a figyelmet!