Nyelvi erőforrások a Szegedi Tudományegyetemen

MTA-SZTE Mesterséges Intelligencia Kutatócsoport
Szegedi Tudományegyetem, Informatikai Intézet

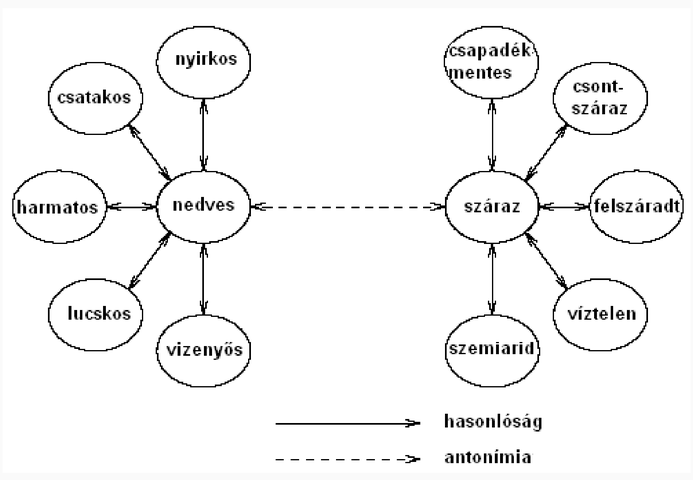
Szeged Korpusz és Treebank

- 82 000 mondat
- 1,5 millió szövegszó
- 230 000 írásjel
- 6 domén
  - iskolai fogalmazások
  - számítógépes szövegek
  - irodalom
  - jogi szövegek
  - újságcikkek
  - üzleti rövidhírek
- Kézzel ellenőrzött morfológiai és szintaktikai (konstituens és függőségi) elemzés
- Névelemek annotációja
- Félig kompozicionális szerkezetek annotációja
- Koreferenciaviszonyok annotációja

http://rgai.inf.u-szeged.hu/SzegedTreebank

Magyar WordNet (HuWN)

- Lexikális adatbázis
- Fogalmak hálóba rendezve különféle lexikai szemantikai relációk alapján
- 40 000 synset (általános ontológia) + 2000 üzleti nyelvi, ill. 650 jogi nyelvi synset



http://rgai.inf.u-szeged.hu/HuWN

Szeged Paralell

- Magyar-angol párhuzamos korpusz
- Kézzel párhuzamosított bekezdés és mondat szinten:
  - Nyelvkönyvek
  - EU-s szövegek
  - Kétnyelvű újságok
  - Irodalom
- 99 000 mondatszintű egység
- Egy része félig kompozicionális szerkezetekre annotálva

http://rgai.inf.u-szeged.hu/corpus_paralell

HunLearner

- Középhaladó és haladó szintű magyarul tanulók fogalmazásai
- Számítógépen, szótár és nyelvkönyv nélkül írt fogalmazások
- 1400 mondat
- Főnévi morfológiai hibák és alanyi/tárgyas ragozási hibák jelölve

Az elején nagyon nehéz volt a szituació. Nem ismertem senkit, nem volt sok pénzem, nagyon hiányozott a családom... Irodakat tisztítottam és nem kaptam elég pénzt, hogy túléljek, mert London nagyon drága város. Pár hét múlva elkezdtem baby szitterkedni (az volt az extra munkám). Most elég pénzem van mindenre, amire szükségem van. Háromszor hazámentem Magyarországra, hogy meglássam a családomat. Minden nap beszélek velük skype-on, úgyhogy már nem hiányoznak annyira, mint az elején. Itt sok új barátom van, akikkel csak angolul beszélek. Néha sájnalom, hogy senkivel nem tudok magyarul beszélni, de ilyen az élet. A magyar baratáim megigerték, hogy Londonba jönnek vendégségbe, de még mindig nem jöttek. Tudom, hogy nagyon drága a jegy, de már egy éve itt lákom, és még mindig nem jöttek. Biztos féltékenyek vannak rám.

http://rgai.inf.u-szeged.hu/hunlearner

Korpuszaink és adatbázisaink ingyenesen hozzáférhetők oktatási és kutatási célokra.

Bizonytalanságra annotált korpuszok

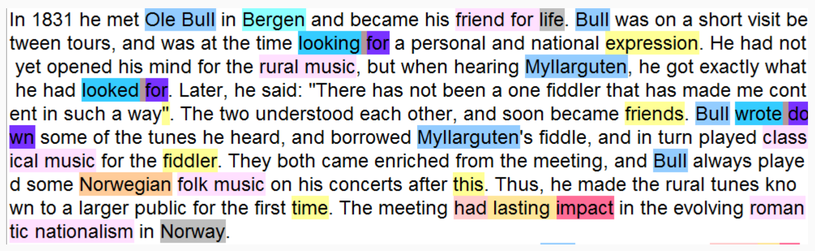
- BioScope (20K mondat)
  - Orvosi szövegek
  - Biológiai absztraktok
  - Biológiai cikkek
- CoNLL-2010 Shared Task korpuszok (biológiai cikkek (18K mondat) + Wikipedia-szócikkek (20K mondat))
- WikiWeasel 2.0: diskurzusszintű bizonytalanság (Wikipedia-szócikkek)
- hUnCertainty: magyar korpusz (20K mondat) (bűnügyi hírek, Wikipédia-szócikkek, és közösségi médiából származó bejegyzések)

De nem biztos hogy mindenkinek telik 1000Ft /fő /nap kajára!
Na ez olyan de mégis más.
Elképzelhető, hogy a második világháborúban elesett katonától származnak a csontok, de az is lehet, hogy a közelben fekvő sírkert régi temetője volt korábban itt.

http://rgai.inf.u-szeged.hu/uncertainty

Többszavas kifejezések korpuszai

- Wiki50 korpusz:
  - 50 angol Wikipedia-szócikk (4700 mondat)
  - MWE-k és NE-k kézzel jelölve
- Szeged Treebankben és SzegedParalell egy részében félig kompozicionális szerkezetek
- JRC-Acquis jogi párhuzamos korpuszban félig kompozicionális szerkezetek angol, német, spanyol és magyar nyelven (100K token minden nyelven)


In 1831 he met Ole Bull in Bergen and became his friend for life. Bull was on a short visit between tours, and was at the time looking for a personal and national expression. He had not yet opened his mind for the rural music, but when hearing Myllarguten, he got exactly what he had looked for. Later, he said: "There has not been a one fiddler that has made me content in such a way". The two understood each other, and soon became friends. Bull wrote down some of the tunes he heard, and borrowed Myllarguten's fiddle, and in turn played classical music for the fiddler. They both came enriched from the meeting, and Bull always played some Norwegian folk music on his concerts after this. Thus, he made the rural tunes known to a larger public for the first time. The meeting had lasting impact in the evolving romantic nationalism in Norway.

http://rgai.inf.u-szeged.hu/mwe

Névelemek korpuszai

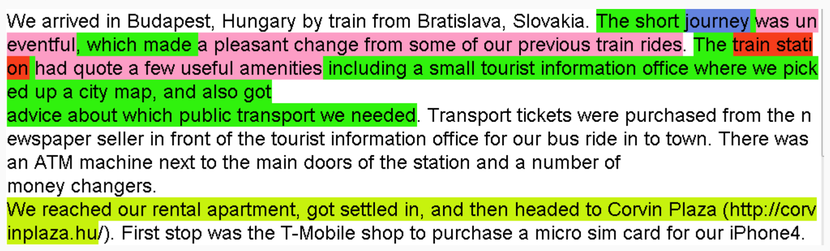
- személy-, hely-, szervezetnevek és egyéb tulajdonnevek
- 220 000 szövegszó (Szeged Korpusz üzleti rövidhírek)
- 470 000 szövegszó (HVG-cikkek)
  - Szó szerinti jelölés
  - Metonimikus jelölés

Elutaztunk Barcelonába.
A Barcelona nyerte a Bajnokok Ligáját.

http://rgai.inf.u-szeged.hu/corpus_ne

SzegedTrip

- 500 utazási blog 5 úticélhoz kapcsolódva
- Angol nyelvű
- Pozitív és negatív vélemények adott dologra vonatkoztatva
- Személyiségjegyekre utaló szövegrészek is jelölve


We arrived in Budapest, Hungary by train from Bratislava, Slovakia. The short journey was uneventful, which made a pleasant change from some of our previous train rides. The train station had quote a few useful amenities including a small tourist information office where we picked up a city map, and also got
advice about which public transport we needed. Transport tickets were purchased from the newspaper seller in front of the tourist information office for our bus ride in to town. There was an ATM machine next to the main doors of the station and a number of
money changers.
We reached our rental apartment, got settled in, and then headed to Corvin Plaza (http://corvinplaza.hu/). First stop was the T-Mobile shop to purchase a micro sim card for our iPhone4.

http://rgai.inf.u-szeged.hu/szegedtrip

Legújabb korpuszaink

- SZEMEK: magyar nyelvű, orvosi témájú újságcikkek korpusza (SZTE ÁOK-val együttműködésben)
- Dívány: termékvélemények szentimentre annotálva (Precognox Kft.-vel együttműködésben)
- Miskolc Jogi Korpusz (Miskolci Egyetemmel együttműködésben)
- PARSEME Shared Task Corpora: igei többszavas kifejezések 20 nyelvre nemzetközi együttműködésben