MTA Sztaki Nyelvtechnológiai Kutatócsoport
http://hlt.sztaki.hu

Hun* Nyelvtechnológiai Eszközök

BME Média Oktató és Kutató Központ
http://mokk.bme.hu

Ács Judit[1], Gyepesi György[2], Halácsy Péter[3], Makrai Márton[1], Nagy Viktor[3], Nemeskey Dávid[1], Németh László[4], Pajkossy Katalin[1], Recski Gábor[1], Rung András[5], Simon Eszter[6], Trón Viktor[7], Varga Dániel[3], Zséder Attila[1], Kornai András[1]

[1]MTA Sztaki [2]Google [3]Prezi [4]OpenOffice.org [5]Ergomania [6]MTA NyTI [7]Nuance

Huntoken

Mondatokra bontás:

- mondathatár alapesetben [.!?] karakter után nagybetűs szó
- rövidítések, címek, dátumok és számok is tartalmazhatnak *pont* karaktert
- rövidítésre is végződhet mondat

Szavakra bontás:

- szóhatár: alapesetben a szóköz
- írasjelek előtt a szóközt elhagyjuk
- speciális karakterek jó kezelése – pl. kötőjel vagy gondolatjel megkülönböztetése
- mondaton belüli idézés kezelése

Tulajdonságok:

- nagy pontosság (98%)
- szóközök normalizálása, gyakori magyar rövidítések, mértékegységes számok, szavakon belüli speciális karakterek kezelése

Hunspell

Helyesírás-ellenőrzés és -javítás:

- bemenet: szavak
- egyik kimenet: helyes/helytelen
- másik kimenet: helytelen esetén javítási javaslatok

Tulajdonságok:

- széles körben elterjedt (LibreOffice, OpenOffice.org, Mozilla Firefox 3+, Thunderbird, Google Chrome)
- érthető, letisztult fájlformátum
- új nyelvek hozzáadása egyszerű
- szótár kb. 100 nyelvre
- gyors a gazdag morfológiájú nyelvekre is, mint a magyar
- információk, letöltés: `http://hunspell.sourceforge.net/`

Hunmorph

Morfológiai elemzés:

- toldalékok (ragok, képzők) felismerése
- *eszed* (1): *ész* főnév E/2. birtokos személyraggal
- *eszed* (2): *eszik* jelen idejű, E/2., kijelentő módú, határozott ragozású igealak
- gyerekeinket: gyerek/NOUN<PLUR><POSS<1><PLUR>><CAS<ACC>>

NOUN	főnév
<PLUR>	többesszám
<POSS<1><PLUR>>	birtokos, T/1.
<CAS<ACC>>	tárgyrag

Tulajdonságok:

- ugyanazokat az erőforrásokat használja, mint a *Hunspell*
- új nyelvek hozzáadása egyszerű, *MorphDB*

Általános információk

- a felsorolt szoftverek szabadon hozzáférhetőek és nyílt forráskódúak
- a megnevezett erőforrások szabadon hozzáférhetőek
- a fel nem tüntetett szoftverek és erőforrások helye: `http://mokk.bme.hu/resources/`
- több mint 100 független, magyar nyelvű hivatkozás
- több mint 600 független hivatkozás összesen

Hunpos

Szófaji elemzés:

- bemenet: tetszőleges szöveg mondatokra és szavakra bontva (pl. *Huntoken*)
- kimenet: minden szó szófajjal ellátva

Szavak	A	kutya	húst	evett
Szófaj	ART	NOUN	NOUN	VERB
Szótő	az	kutya	hús	eszik
Morf. elemzés	ART	NOUN	NOUN<CAS<ACC>>	VERB<PAST>

Tulajdonságok:

- nagy pontosság (97% fölött)
- figyelembe veszi a szövegkörnyezetet (*eszed* mint főnév vagy ige)
- általánosan felhasználható: új erőforrásokkal tetszőleges címkehalmazt egyértelműsít
  - morfológiai jegyeket – *Hunmorph*
    - eszed 1: ész(NOUN<POSS<2>>)
    - eszed 2: eszik(VERB<PERS<2>><DEF>)
- egy nagyságrenddel gyorsabb, mint a hasonló pontosságú, komplexebb rendszerek
- nagy pontosság a még nem látott szavak esetén
- gyors a gazdag morfológiájú nyelvekre is, mint a magyar
- információk, letöltés: Google Code

HunTag

Szekvenciális címkéző

- többszavas kifejezések felismerése:
  - tulajdonnevek: *HunNer*
  - mondattani frázisok: *HunChunk*
- információk, letöltés: `https://github.com/recski/HunTag/`
- könnyen adaptálható új feladatra, új nyelvekre
- moduláris –> könnyen bővíthető
  - új jellemzőkkel (*jegyekkel*), pl. nagybetűs; névelő van előtte
  - külső erőforrásokkal, pl. névlisták

HunNer

- tulajdonnevek felismerése strukturálatlan szövegben
- magyar és angol nyelvre
- kategóriák: személy, hely, szervezet, egyéb
- magyarra a legjobb publikált pontosság (96.1)
- angolra a legjobb publikálthoz hasonló eredmény (86.3)

HunChunk

- *chunk*-ok: a mondat legnagyobb funkcionális egységei
- pl. *a mérges ember harapós kutyája*
- magyarra a legjobb 81.7
- angolra hasonló eredmény (75.0), mint a legjobbak

Hunalign

Mondatok párhuzamosítása:

- bemenet: kétnyelvű, mondatokra bontott szöveg
- kimenet: egymásnak megfelelő mondatpárok (nyelvenként akár több mondat)
- kihívások:
  - szabad fordítások az adatban
  - néha egy mondatot egy másik nyelvre nem fordítanak, vagy
  - több mondatban fordítanak
- gépi fordító eszközök bemenetéül szolgál

Tulajdonságok:

- egyszerű, gyors
- sok ingyenesen elérhető, párhuzamos szöveg előállítására használják
  - *Hunglish*
  - JRC-Acquis Corpus: az egyik leggyakrabban használt, többnyelvű szöveges adat

Magyar webkorpusz

- első magyar gigaword (legalább 1Mrd szavas) korpusz (szövegyűjtemény)
- teljes méret: 3.5M weboldal, 1.48Mrd szó
- 2003-as web alapján
- helyesírás-ellenőrzés alapján (*Hunspell*) több elérhető vezió:
  - teljes adat
  - < 40% ismeretlen szó – nem magyar nyelvű dokumentumok szűrése
  - < 8% ismeretlen szó – az ékezet nélküli és egyéb problémás oldalak szűrése
  - < 4% ismeretlen szó – nyomtatásban megjelent dokumentumokhoz hasonló minőségű
- gyakorisági szótár: `http://szotar.mokk.bme.hu/szoszablya/searchq.php`
- ezek közül a legjobban megszűrt változat érhető el a fenti linken - 4GB

Mire jó?

- nyelvészeti kutatásokhoz
- pszicholingvisztikai kísérletek tervezéséhez

MorphDB

Morfológiai adatbázis

- a magyar nyelv legteljesebb, nyilvánosan elérhető, elméletileg megalapozott morfológiai leírása: morphdb.hu
- angolra: morphdb.en
- három létező lexikon egyesítésével készült, melyek különböző elméletek különböző kódolásait tartalmazzák
- összegzi és egységesíti az eddigi elméleteket és szótárakat
  - magyar ISpell szótár
  - Elekfi László ragozási szótára
  - Füredi–Kelemen-féle szépprózai gyakorisági szótár
  - Papp Ferenc szóvégmutató szótára

Szófaj	főnév	melléknév	ige	határozószó	egyéb (10 kat.)
Szótövek	88.026	17.514	12.549	1.932	1.547

Felhasználás:

- *Hunmorph*
- magyarlanc: szintaktikai elemző és szófaji egyértelműsítő

Hunglish

angol-magyar párhuzamos szöveg

- legnagyobb angol-magyar párhuzamos szöveg
- méret: 4M mondat, 115M szó
- heterogén: szépirodalom, jogi szövegek, filmfeliratok, szoftver dokumentációk
- webes felületen kereshető és egészben letölthető
- teljesen automatizált szövegpárhuzamosítás
- Mire jó?
  - gépi fordítás
  - nyelvtanulás (szövegkörnyezetek, különböző jelentések)
- információk, letöltés, keresés: `http://hunglish.hu`

Hunglish 2.0

- több mondatpár
- megújult webes felület
  - dokumentumpárok feltöltése többféle formátumban
  - jó/rossz fordítások felhasználói értékelése
  - elgépelések, helyesírási hibák javítása
- duplikátumszűrés

Felhasználói statisztikák

- 150K kérés 10K felhasználónak havonta
- 2000 bejelentett hibás fordítás
- 37 új dokumentum hozzáadása, 17K új mondatpár