A Hunglish Korpusz és szótár

Halácsy Péter[1], Kornai András[1], Németh László[1], Sass Bálint [2] Varga Dániel[1], Váradi Tamás[2] Vonyó Attila

[1] BME – Média Oktató és Kutató Központ
1111 Budapest, Stoczek u. 2
{hp,nemeth,daniel}@mokk.bme.hu
[2] MTA Nyelvtudományi Intézet
1068 Budapest, Benczúr u. 33.
{joker,varadi}@nytud.hu

MSZNY - 2005. december 8.

Eredmények

- Hunglish Korpusz: mondatszinten illesztett magyar–angol párhuzamos korpusz
- hunalign: mondatszintű illesztő párhuzamos korpuszok építéséhez
- Steinbeck Korpusz: manuálisan illesztett párhuzamos szöveg
- szótár: párhuzamos gyakorisági adatokkal bővített angol–magyar szótár
- kereső: webes keresőrendszer párhuzamos korpuszokhoz

Nyersanyag forrásai

A gyakran hivatkozott automatikus módszerek (Resnik 2002) helyett, manuálisan gyűjtöttünk párhuzamos szövegeket, elsősorban az internetről.

- Irodalmi szövegek. Fő forrásunk a *Project Gutenberg* és a *Magyar Elektronikus Könyvtár*.
- Jogi szövegek. Az EU közösségi jogszabályok *CELEX adatbázisa* és az *Európai Alkotmány*.
- Nyílt forráskódú szoftverek dokumentációi. KDE, Gnome, OpenOffice, Mozilla és GNU.
- Filmfeliratok. Az internetről letölthető jogvédett szövegek.
- Magazinok angol és magyar kiadásai.
- Sajtófigyelő. A Magyar Telekom Rt. kétnyelvű sajtófigyelő adatbázisa.

A korpusz összetétele szövegtípusok szerint

forrás	**Angol tokenek (m)**	**Magyar tokenek (m)**
irodalom	14.6	11.5
jogi	24.1	18.3
filmfelirat	2.5	1.9
szoftver	0.8	0.7
magazinok	0.3	0.3
sajtó	2.1	1.7
összesen	**44.5**	**34.5**

Szerzői jogi kérdések

A jogvédett párhuzamosított szövegeket mondatszintű keverésnek vetettük alá.

- A legfontosabb alkalmazásaink számára nem jelent hátrányt, beleértve statisztikus gépi fordítórendszerek tanítását is.
- Lehetetlenné teszi nagyobb szövegrészek rekonstruálását, védve a szerzői jogok tulajdonosainak érdekeit.

A Hunglish Korpuszt auditálta és hamarosan terjeszteni fogja a Linguistic Data Consortium.

hunalign mondatpárhuzamosító

- Nagy pontosságú és fedésű.
- Nyelvfüggetlen.
- Hatékonyan képes hasznosítani kétnyelvű szótárat és szótövezőket, de erőforrás-mentesen is pontos.
- Gyors.

Az algoritmus

bemenet

1. Mondatra darabolt forrás- és célnyelvi szöveg.
2. Kétnyelvű frázislexikon.

1. Egyszerű nyersfordítás építése a forrásszövegből, a célnyelvi gyakoriságok figyelembe vételével.
2. Első párhuzamosítás, a nyersfordítás és mondathossz-hasonlóság figyelembe vételével.
3. Lexikon bővítése automatikus szótárépítő eljárással.
4. Megismételt párhuzamosítás, a bővített lexikon felhasználásával.

Algoritmus értékelése

Eljárás	**pontosság**	**fedés**
len	97.58	97.55
len+id	97.65	97.42
dic	97.30	97.08
len+dic	98.86	98.88
len+dic+stem	99.34	99.34
len+boot	98.63	98.74
len+boot+stem	99.12	99.18

id - szóazonosság
len - karakterszám
dic - kétnyelvű lexikon
boot - automatikus lexikonbővítés
stem - szótövező alkalmazása

Algoritmus értékelése

Eljárás	**pontosság**	**fedés**
len	97.58	97.55
len+id	97.65	97.42
dic	97.30	97.08
len+dic	98.86	98.88
len+dic+stem	99.34	99.34
len+boot	98.63	98.74
len+boot+stem	99.12	99.18

id - szóazonosság
len - karakterszám
dic - kétnyelvű lexikon
boot - automatikus lexikonbővítés
stem - szótövező alkalmazása

- A klasszikus Gale-Church algoritmus.

Algoritmus értékelése

Eljárás	**pontosság**	**fedés**
len	97.58	97.55
len+id	97.65	97.42
dic	97.30	97.08
len+dic	98.86	98.88
len+dic+stem	99.34	99.34
len+boot	98.63	98.74
len+boot+stem	99.12	99.18

id - szóazonosság
len - karakterszám
dic - kétnyelvű lexikon
boot - automatikus lexikonbővítés
stem - szótövező alkalmazása

- A klasszikus Gale-Church algoritmus.
- Legjobb eredményünk nyelvi erőforrások alkalmazásával.

Algoritmus értékelése

Eljárás	pontosság	fedés
len	97.58	97.55
len+id	97.65	97.42
dic	97.30	97.08
len+dic	98.86	98.88
len+dic+stem	99.34	99.34
len+boot	98.63	98.74
len+boot+stem	99.12	99.18

id - szóazonosság
len - karakterszám
dic - kétnyelvű lexikon
boot - automatikus lexikonbővítés
stem - szótövező alkalmazása

- A klasszikus Gale-Church algoritmus.
- Legjobb eredményünk nyelvi erőforrások alkalmazásával.
- Legjobb eredményünk nyelvi erőforrások alkalmazása nélkül.

hunalign erőforrás nélkül

A `hunalign` és (Moore 2002) összehasonlítása három szövegen, csak az egy az egyhez szegmentumokon.

feladat	`hunalign`		`Moore '02`	
	pont.	**fed.**	**pont.**	**fed.**
1984 Hun-Eng tövezett	99.22	99.24	99.42	98.56
1984 Hun-Eng nem töv.	98.88	99.05	99.24	97.39
1984 Rom-Eng nem töv.	97.10	97.98	97.55	96.14
Cup of Gold Hun-Eng töv.	97.03	98.44	96.45	97.53

Nyelvfüggetlenség

A `hunalign` pontossága és fedése a MULTEXT-East 1984 korpuszon különböző angol–X nyelvpárokra, nyelvi erőforrások használata nélkül.

nyelv	**pontosság**	**fedés**
észt	99.34	99.53
cseh	98.60	98.75
román	97.10	97.98
szlovén	99.44	99.61

Steinbeck korpusz

John Steinbeck Egy marék arany című művének manuálisan illesztett változata.

- Nyelvenként körülbelül 230 oldal, 5400 mondat, 57,000 szó.
- 6 emberhétnyi manuális munka.
- Elsősorban mondatpárhuzamosítás hatékonyságának mérésére szolgál.
- Csak kutatási célra használható fel.

Hunglish szótár

Vonyó Attila ismert szótárából kiindulva, azt

- együtt-előfordulási statisztikákkal láttuk el a morfológiailag elemzett Hunglish Korpusz alapján.
- a Hunglish korpuszon végzett automatikus szótárépítés eredményével bővítettük.

Statisztikus gépi nyersfordítás céljaira építettük, de későbbi alapja lehet szótár-szolgáltatásnak is.

Keresőfelület

Párhuzamos korpuszokban való keresést tesz lehetővé:

- szótőre vagy teljes szóalakra.
- szavakkal vagy kifejezésekkel.
- logikai operátorokkal.
- akár mindkét nyelvre.

Keresőfelület

Magyar Tudományos Akadémia
Nyelvtudományi Intézet

This is a beta test of our sentence search. The sentences are translations of each other. Use this t ool like a bilingual lexicon. The sentences have been aligned automatically. We use the hunmorph morphological analyzer to find roots of the words.

please post your comments to the open forum

Hungarian: akaratom ellenére English: all reset search

Results 1 - 20 of 5639 from the Hunglish corpus

Movie subtitles	- **Akaratom ellenére** visznek.	- They're taking me against my will!
Literature	**Akaratom ellenére** hoztak ide.	I was brought here against my will."
Literature	Lehet, **akaratom ellenére** is...	Possibly against my own will..."
Literature	Tekintetem **akaratom ellenére** találkozott a lányéval.	Before I could stop myself I'd caught the girl's eye.
Literature	Így történt, hogy **akaratom ellenére** Wolf Larsen szolgálatába léptem.	And thus it was that I passed into a state of involuntary servitude to Wolf Larsen.
Literature	**Akaratom ellenére** vágtunk át Mória homályán, s lám, kárát vallottuk.	Against my will we passed under the shades of Moria, to our loss.
Literature	**Akarata ellenére**	He couldn't help it.
Literature	Senki sem varrhatja magát a nyakamba az **akaratom ellenére**, hát nyugi.	But nobody imposes on me against my will, so relax about it.

Keresőfelület

Hungarian: English: +stroke.of luck Literature reset search

Results 1 - 20 of 70 from the Literature collection of the Hunglish corpus

Literature	Cathcart ezredes a jó szerencsének éppen ilyen csapásáért imádkozott, mint Duluth őrnagy halála.	Colonel Cathcart had been praying for just some **stroke of** good **luck** like Major Duluth's death.
Literature	És egy év elteltével a szerencse úgy hozta, hogy valóban kaptam.	And sure enough, after about a year, by a **stroke of luck** it happened.
Literature	- Még szerencse, hogy időben megtudtam a dolgot - folytatta Mr. Diggory feje.	"- it's a real **stroke of luck** I heard about it," said Mr. Diggory's head.
Literature	Azt mondom, a szerencse hozta úgy, de az igazat megvallva, úgyis talpra álltam volna.	I say by a **stroke of luck**, but the fact is that I was bound to fall on my feet.
Books of Arthur C. Clark	Ráadásul a zseniális szerkesztő az eredeti Kilátás az Olümposzról címet az Isteni szenvedélyekre változtatta.	And by a **stroke of luck**, an editor **of** genius had changed her original title, The View from Olympus, to The Passions **of** the Gods.
Literature	Éjfélt ütött!)	**Stroke of** midnight!)
Literature	Zseniális ötlet!	"It's a **stroke of** genius!"
Literature	Ezt nevezem én szerencsének.	A pretty **stroke of** fortune!
Literature	- Kevés talpraesettebb ötlete támadt.	It was a **stroke of** insight.
Literature	Csak várom, hogy lesújtson a végzet.	I wait for some **stroke of** doom.'
Literature	Ez a végzet utolsó csapása!	It is the last **stroke of** doom!'

Összefoglalás

- A korpuszok és erőforrások mellett eszközeinket is publikáltuk.
- A korpuszt már használják is többen.
- Módszereink nyelvfüggetlenek, megismételhetők más nyelvpárokra is.

Linkek

Hunglish Korpusz

http://mokk.bme.hu/eszkozok/hunglishkorpusz

hunalign

http://mokk.bme.hu/eszkozok/hunalign

Kereső

http://hunglish.hu