

NYELVTECHNOLÓGIA

Sass Bálint

sass@digitus.itk.ppke.hu

Pázmány Nap
2007. október 17.

1 MI AZ?

2 TÖBBÉRTELMŰSÉG

3 KUTATÁS

1 MI AZ?

2 TÖBBÉRTELMŰSÉG

3 KUTATÁS

BEVEZETŐ

„Language makes us human.”

Turing–teszt

nyelv és gondolkodás

- nyelv → gondolkodás (Sapir–Whorf)
- minden nyelv azonos mértékben alkalmas a gondolatok kifejezésére

nyelv, nyelvi képesség vizsgálata

→ agy működése, intelligencia

MI A NYELVTECHNOLÓGIA?

- „Célja a beszélő gép létrehozása.”
szükséges hozzá: beszédértés és -szintézis, mély megértés, információ kivonatolás, következtetés
- „Az emberi nyelvhasználatot leíró modellek gyakorlati, számítógépes megvalósításával foglalkozik.”
- „Az emberi nyelvi képesség számítógépes aspektusaival foglalkozik.”
- „Az információfeldolgozásnak a természetes nyelvek kezelésére szakosodott területe.”
- „Az internet nagyrészt szöveg. Nyerjük ki a szükséges/lényeges tartalmat.”
- „A nyelvi technológiák szerepe kettős: egyfelől rendkívül fontos eszközei az európai kulturális örökség megőrzésének, másfelől a gazdasági növekedés forrását jelentik.” (EU)

KAPCSOLÓDÓ TUDOMÁNYOK

- nyelvészet, pszichológia, logika
- kognitív tudomány
- matematika, statisztika
- számítástudomány, mesterséges intelligencia
- szoftvertechnológia, adatbányászat

TERÜLETEI

Elemzési szintek:

- morfológia
- szintaxis
- szemantika: WSD, ontológiák, anafora-feloldás ...

Alkalmazások:

- helyesírásellenőrzés, elválasztás
- kereséstámogatás
- információ kivonatolás és osztályozás
- természetes nyelvi interfészek
- gépi fordítás

SZABÁLYOK VS. STATISZTIKA

számítógépek teljesítménynövekedése + korpusz alapú gépi tanuló eljárások kifejlesztése → ugrásszerű fejlődés

korpusz: elemzett szöveggyűjtemény

Magyar Nemzeti Szövegtár

<http://corpus.nytud.hu/mnsz>

Magyar Webkorpusz

<http://mokk.bme.hu/resources/webcorpus>

1 MI AZ?

2 TÖBBÉRTELMŰSÉG

3 KUTATÁS

TÖBBÉRTELMŰSÉG: MORFOLÓGIA

csontváza

darabot

középállás

<http://corpus.nytud.hu/~joker/demo/egy>

TÖBBÉRTELMŰSÉG: MORFOLÓGIA

csont | váz | a

darab | o | t

közép | állás

<http://corpus.nytud.hu/~joker/demo/egy>

TÖBBÉRTELMŰSÉG: MORFOLÓGIA

csont | váza

dara | bot

közé | pállás

<http://corpus.nytud.hu/~joker/demo/egy>

TÖBBÉRTELMŰSÉG: „MÉLY MEGÉRTÉS”

„Ő az, aki három éve beugrott helyettesíteni a leányfalui lelkiyakorlatos házba előadóként meghívott, de időközben megsebesült Cantalamessa atyát, a pápa gyóntatóját.”

„Ha nem vigyázunk, úgy múlik el József Attila centenáriuma, hogy a magyarság nagyobb része nem is értesül róla. Készült ugyan egy filmsorozat a költő életéről és poéziséről, a négyrészes Eszmélet után; ezt 2004 decemberében a Duna Televízió be is mutatta . . .”

TÖBBÉRTELMŰSÉG: „MÉLY MEGÉRTÉS”

„Ő az, aki három éve beugrott helyettesíteni a leányfalui lelkiyakorlatos házba | előadóként meghívott, de időközben megsebesült Cantalamessa atyát, a pápa gyóntatóját.”

„Ha nem vigyázunk, úgy múlik el József Attila centenáriuma, hogy a magyarság nagyobb része nem is értesül róla. Készült ugyan egy filmsorozat a költő életéről és poéziséről, a négyrészes Eszmélet után; ezt 2004 decemberében a Duna Televízió be is mutatta . . . ”

TÖBBÉRTELMŰSÉG: „MÉLY MEGÉRTÉS”

„Ő az, aki három éve beugrott helyettesíteni a leányfalui lelkiyakorlatos házba | előadóként meghívott, de időközben megsebesült Cantalamessa atyát, a pápa gyóntatóját.”

„Ha nem vigyázunk, úgy múlik el József Attila centenáriuma, hogy a magyarság nagyobb része nem is értesül róla. Készült ugyan egy filmsorozat a költő életéről és poéziséről, a négyrészes *Eszmélet* után; ezt 2004 decemberében a Duna Televízió be is mutatta . . .”

TÖBBÉRTELMŰSÉG: „MÉLY MEGÉRTÉS”

„Ő az, aki három éve beugrott helyettesíteni a leányfalui lelkiyakorlatos házba | előadóként meghívott, de időközben megsebesült Cantalamessa atyát, a pápa gyóntatóját.”

„Ha nem vigyázunk, úgy múlik el József Attila centenáriuma, hogy a magyarság nagyobb része nem is értesül róla. Készült ugyan egy filmsorozat a költő életéről és poéziséről, a négyrészes *Eszmélet után*; ezt 2004 decemberében a Duna Televízió be is mutatta . . . ”

1 MI AZ?

2 TÖBBÉRTELMŰSÉG

3 KUTATÁS

TÉMÁM: MAGYAR IGEI KERETEK

„Altschuler elégedetlenül ráncolta össze a homlokát, egyesek semmiféle kapcsolatban sem álltak az ellenzéki mozgalommal, nyilván csak azért kerültek be a névsorba, mert Rákosi egy füst alatt más természetű vitáit is rendezni kívánta, de a valóban résztvevők is csak a második-harmadik vonulatát alkották a mozgalomnak.”

TÉMÁM: MAGYAR IGEI KERETEK

„Altschuler elégedetlenül ráncolta össze a homlokát, egyesek semmiféle kapcsolatban sem álltak az ellenzéki mozgalommal, nyilván csak azért kerültek be a névsorba, mert Rákosi egy füst alatt más természetű vitáit is rendezni kívánta, de a valóban résztvevők is csak a második-harmadik vonulatát alkották a mozgalomnak.”

- összeráncol homlokát
- áll kapcsolatban vmivel
- bekerül vmibe
- rendez vmit (füst alatt)
- alkot vmijét vminek

Célom: a magyar igék bővítményszerkezetének feltárása.

KERET-BÖNGÉSZŐ: „MAZSOLA”

Mazsola - a magyar igei bővítményszerkezet vizsgálata - Mozilla Firefox

http://carnet.egi-bin.mazsola.mazsola_hun.pl

Kopiar: teljes MWSZ

Először:

Igék: jésk

Nem: Esetnévűlő: [] be Nem: Vonzatló: []

Nem: Esetnévűlő: [] Nem: Vonzatló: []

Nem: String: []

Méret

4618 találat. [fogság](#) [412] [pánik](#) [320] [lat](#) [208] [teher](#) [318] [kétség](#) [206] [csapda](#) [145] [kút](#) [164] [hiba](#) [104] [túlzás](#) [114] [bűn](#) [137] [hadifogság](#) [68] [kóma](#) [66] [kísértés](#) [66] [kategória](#) [61] [gondolkodó](#) [58] [út](#) [110] [szerelem](#) [60] [véglet](#) [38] [verem](#) [28] [víz](#) [60] [tévédés](#) [40] [depresszió](#) [33] [transz](#) [23] [nyak](#) [41] [tartomány](#) [36] [zavar](#) [33] [önkívület](#) [18] [búskomorság](#) [16] [extázis](#) [17] [másik](#) [49] [gódor](#) [30] [rabság](#) [17] [irány](#) [27] [betegség](#) [27] [vétek](#) [16] [révület](#) [13] [ügy](#) [11] [késedelem](#) [16] [büvölés](#) [13] [kör](#) [20] [Duna](#) [18] [ármutat](#) [12] [eksztrázis](#) [10] [te](#) [22] [vonal](#) [18] [tenger](#) [18] [sáv](#) [18] [pár](#) [14] [kelepcse](#) [9] [tergár](#) [8] [szarvas](#) [8] [sárga](#) [8] [sódor](#) [8] [kút](#) [13] [falakör](#) [16] [saj](#) [15] [szem](#) [12] [szent](#) [8] [szél](#) [8] [szélke](#) [7] [szelvény](#) [7] [szél](#) [8] [szél](#) [8] [szél](#) [7] [szél](#) [14] [szél](#) [8] [szél](#) [7] [szél](#) [14] [szél](#) [8] [szél](#) [7] [szél](#) [8] [szél](#) [14] [szél](#) [8]

ablakörög
 a hogy a szándékos alkalmatosság az ezerm.
 ezerm.
 az a nemzetállalás jövőké tekintetében meghatározott 25 százalékosan megadható aktívabb értele, megadható a jövőké.
 hogy hi milyen aktívabb értele, az már egyenértékű változó.
 szent
 hogy az értele szent emi Chandra (idén december 14-én kezdődik), a Fény Szerepe.
 aggodalom
 Ahhoz a belgyógyász újrat aggodalomba értele.
 Ahhoz a belgyógyász újrat aggodalomba értele.
 Döntési az szent aggodalomba értele a kettős értele.
 aggodalom
 hogy a hi az aggodalma értele.
 agya
 meg bele értele az agyba.
 szent
 Ezik lefele valóan aktívabb.
 alj
 a hátsó autóbólak aljba értele.
 allergia
 hogy jövőké, depresszió, hányingere, hányingere allergia értele?
 alternatíva
 De rem, meg szent az, meg itt az ezerm értele a nemzetállalás jövőké - szent hiszen szent az is mindig Fennálló - patetikus alternatívákba.

Képz

KERET-BÖNGÉSZŐ: „MAZSOLA”

Mazsola - a magyar igei bővítményszerkezet vizsgálata - Mozilla Firefox

http://carnet.tci.bim.mazsola.mazsola_hun.pl

Kopasz: teljes MNSZ

Ígét: íésk

Nem: Esetnévűlő: [be] Nem: Vonzatt: []

Nem: Esetnévűlő: [] Nem: Vonzatt: []

Nem: String: []

Melhet

4618 találat. [fogság](#) [412] [pánik](#) [320] [lat](#) [298] [teher](#) [318] [kétség](#) [296] [csapda](#) [185] [kút](#) [164] [hiba](#) [164] [túlzás](#) [114] [bűn](#) [137] [hadifogság](#) [88] [kóma](#) [86] [kisértés](#) [66] [kategória](#) [61] [gondolkodó](#) [58] [út](#) [110] [szerelem](#) [56] [véglet](#) [38] [verem](#) [28] [víz](#) [60] [tévedés](#) [40] [depresszió](#) [33] [transz](#) [23] [nyak](#) [41] [tartomány](#) [36] [zavar](#) [33] [önkívület](#) [18] [búskomorság](#) [16] [extázis](#) [17] [másik](#) [49] [gódor](#) [30] [rabság](#) [17] [irány](#) [27] [betegség](#) [27] [vétek](#) [16] [révület](#) [13] [ügy](#) [18] [késedelem](#) [16] [bűvöllet](#) [13] [kór](#) [20] [Duna](#) [18] [ármutat](#) [14] [éksztűzés](#) [10] [te](#) [22] [vonat](#) [16] [tenger](#) [16] [sáv](#) [15] [pár](#) [14] [kelepcse](#) [9] [letargia](#) [4] [szarvas](#) [16] [sódórt](#) [9] [kút](#) [13] [falakör](#) [16] [szó](#) [15] [szem](#) [12] [csont](#) [4] [sz](#) [6] [vénkő](#) [7] [nacsokfennér](#) [7] [in](#) [6] [toto](#) [6] [szé](#) [4] [tete](#) [7] [szé](#) [14] [kőzet](#) [4] [szé](#) [7] [sz](#) [4] [szé](#) [6] [szé](#) [6] [szé](#) [6] [szé](#) [6]

ablakozás
a hogy a szerződési allokációra az eszem,
adnak
és a nemzetállalói járulékos tevékenységben meghatározott 25 százalékosnál magasabb aktívára vonatkozó, magasabb a jövedelmű.
hogy hi milyen aktívára esik, az már egyszerűen változik, az
elvet
hogy adhatna szokatlan Chandra (idén december 14-én kezdődik), a Fény Összege.
aggodalom
Mikor a belgyógyász újratel aggodalomba esett,
Mikor a belgyógyász újratel aggodalomba esett,
Délregei az minijárt megadalmazt esett a kassza elszá.
aggyakut
hogy a hi az aggyakutla esett.
ajta
mag bels esztel az ajtóba.
szé
Ezt lefele valami aktívára.
aj
a hátsó autóbólak aljába esik.
allergia
hogy jövedelmű, depressziós, hányingere allergiába esik?
alternatíva
De rem, mag esztel az, mag itt az eszem esztel a nemzetállalós felállítás - eszt hiszen esztél is mindig Fennálló - patetikus alternatívákba.

Képz

<http://corpus.nytud.hu/mazsola>

Felhasználói név: pazmanynap – jelszó: 2007

ALKALMAZÁS: GÉPI FORDÍTÁS

The screenshot shows the webforditas.hu website in a Mozilla Firefox browser. The page features a navigation bar with logos for 'nemzet Akadémiai Nagyzsúr', 'Akadémiai MoBiMouse', and 'MORPHOLOGIC'. The main content area has a search bar with the text 'angol - magyar' and '197 712 fordítás'. Below the search bar, there is a text area with the following text: 'The thief was caught. The stroke strngs his eye. They talked. He grew into a comely young man. Into a visitor we were added to you.' To the left of the main content, there are three product advertisements: 'MorphoLogic', 'MorphoWord', and 'MoBiCat'. To the right, there is a 'Google Hírek' section with several news items. At the bottom of the page, it says 'powered by Morpho MoBiCat' and 'Design by'. The page number '162' is visible in the bottom left corner.

ALKALMAZÁS: GÉPI FORDÍTÁS

The screenshot shows the webforditas.hu website in a Mozilla Firefox browser. The page features the MorphoLogic logo and navigation links for 'webforditas.hu', 'Online Fordítószoftverek', 'MorphoLogic', and 'Mozilla Firefox'. The main content area includes a search bar with the text 'angol - magyar' and '197.712 fordítás'. Below the search bar, a translation example is shown: 'The thief was caught. The stroke stings his eye. They talked, he grew into a comely young man. Into a visitor we were added to you.' The website also has a sidebar with 'Ajánlott' (Recommended) links to 'MorphoWord' and 'MoBiCat', and a 'Google Hírek' (Google News) section with various news items. The footer contains the text 'Készítette: MofaMorpho MoBiDio' and 'Design by: MofaMorpho MoBiDio'.

<http://www.webforditas.hu>

ALKALMAZÁS: GÉPI FORDÍTÁS

Fülön csípték a tolvajt.

Csípi a szememet a füst.

Kudarcot vallottak.

Egy fess fiataleberré nőtte ki magát.

De: Látogatóba jöttünk hozzátok.

ALKALMAZÁS: GÉPI FORDÍTÁS

Fülön csípték a tolvajt.

→ The thief was caught.

Csípi a szememet a füst.

→ The smoke stings my eye.

Kudarcot vallottak.

→ They failed.

Egy fess fiatalemberré nőtte ki magát.

→ He grew into a comely young man.

De: Látogatóba jöttünk hozzátok.

→ Into a visitor we were added to you.

IGEPROFILOK AUTOMATIKUS ELŐÁLLÍTÁSA

Cél: nyelvfüggetlen (!) igekeret-feltérképező algoritmus
Lehetséges alkalmazás: automatikus szótárgenerálás

forog	20%
forog vmiben	7%
forog vmin	5%
forog vmi körül	4%
forog veszélyben	4%
forog kockán	4%
forog vmivel	2%
forog sírjában	2%
forog vmit	2%
forog veszélyben élete	1%

IGEPROFILOK AUTOMATIKUS ELŐÁLLÍTÁSA

Cél: nyelvfüggetlen (!) igekeret-feltérképező algoritmus
Lehetséges alkalmazás: automatikus szótárgenerálás

húz vmit	17%
húz	5%
húz vmit vmire	3%
húz időt	3%
húz hasznot vmiből	3%

IGEPROFILOK AUTOMATIKUS ELŐÁLLÍTÁSA

Cél: nyelvfüggetlen (!) igekeret-feltérképező algoritmus
Lehetséges alkalmazás: automatikus szótárgenerálás

megköszörül torkát	89%
megköszörül vmit	9%

„ÉLŐSÉG”

Alszik.

Elromlott.

„ÉLŐSÉG”

Alszik.

→ **He** is sleeping.

Elromlott.

→ **It** has gone wrong.

„ÉLŐSÉG”

Alszik.

→ **He** is sleeping.

Elromlott.

→ **It** has gone wrong.

Adott ige alanya élő vagy élettelen?

„ÉLŐSÉG”

Alszik.

→ **He** is sleeping.

Elromlott.

→ **It** has gone wrong.

Adott ige alanya élő vagy élettelen?

Előzetes eredmény – pontosság: 94%

„ÉLŐSÉG”

Alszik.

→ **He** is sleeping.

Elromlott.

→ **It** has gone wrong.

Adott ige alanya élő vagy élettelen?

Előzetes eredmény – pontosság: 94%

Köszönöm a figyelmet!