NYELVTECHNOLÓGIA

Sass Bálint
sass@digitus.itk.ppke.hu

Pázmány Nap
2007. október 17.

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 + korpuszalapú gépi tanulóeljárások kifejlesztése → ugrásszerű fejlődés

korpusz: elemzett szöveggyűjtemény

Magyar Nemzeti Szövegtár

`http://corpus.nytud.hu/mnsz`

Magyar Webkorpusz

`http://mokk.bme.hu/resources/webcorpus`

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 lelkigyakorlatos 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 lelkigyakorlatos 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 lelkigyakorlatos 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 lelkigyakorlatos 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 . . .”

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.”

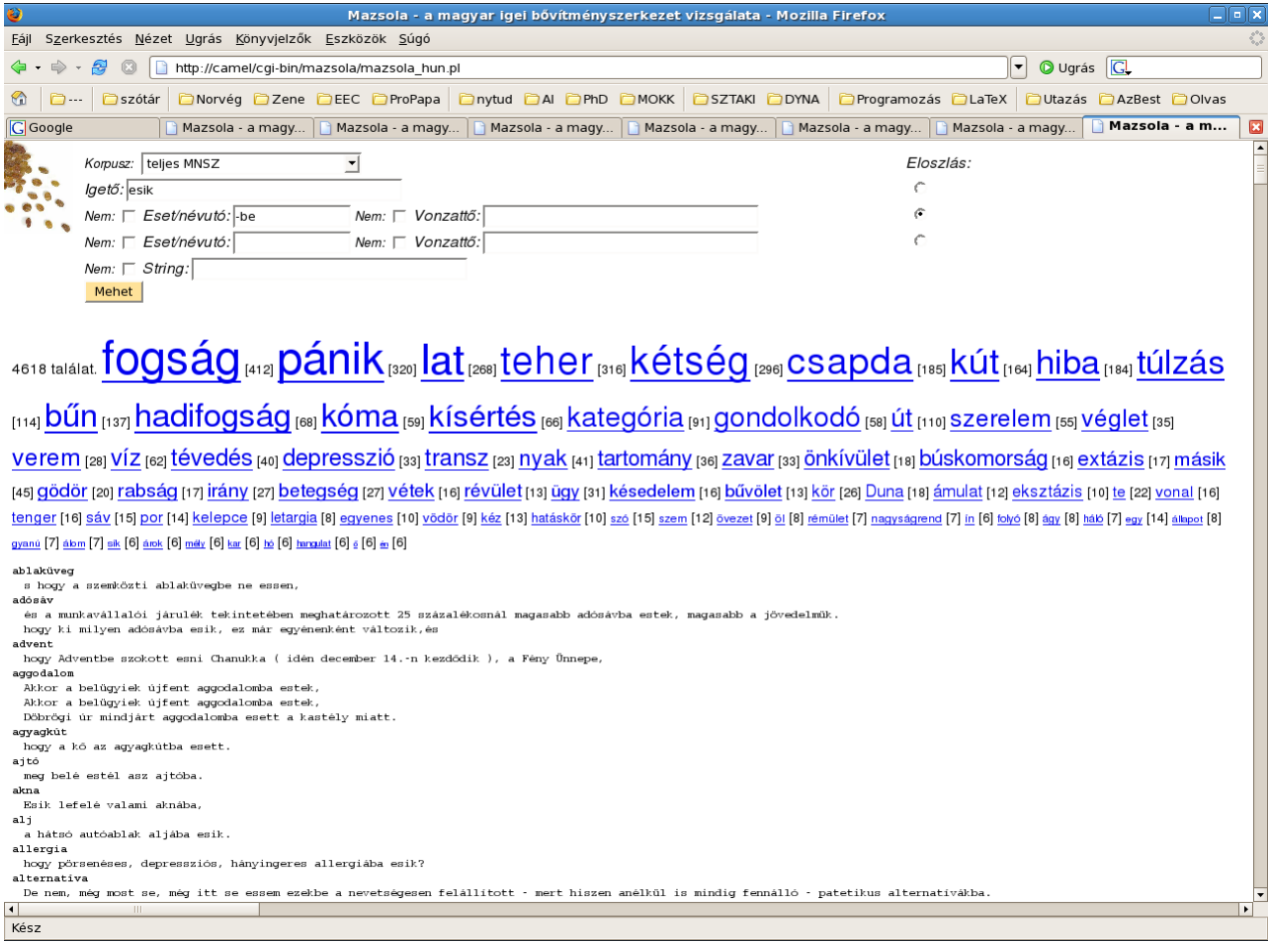
TÉMÁM: MAGYAR IGEI KERETEK

„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.”

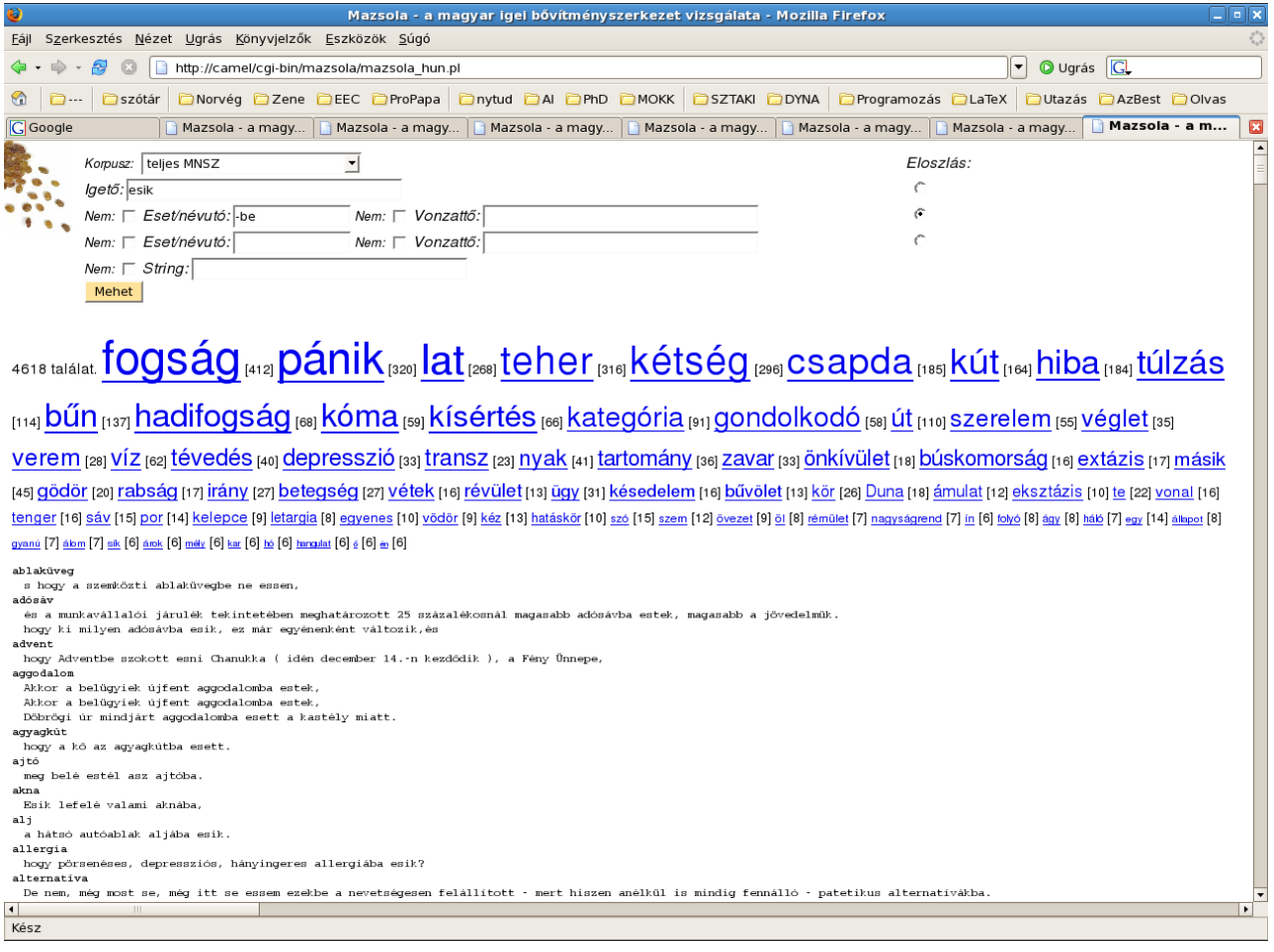
- összeráncol homlokát
- áll kapcsolatban vmivel
- bekerül vmibe
- rendez vmit (füst alatt)
- alkot vmijét vminek

Célom: a magyar igék bővítményszerkezetének feltárása.

KERET-BÖNGÉSZŐ: „MAZSOLA”



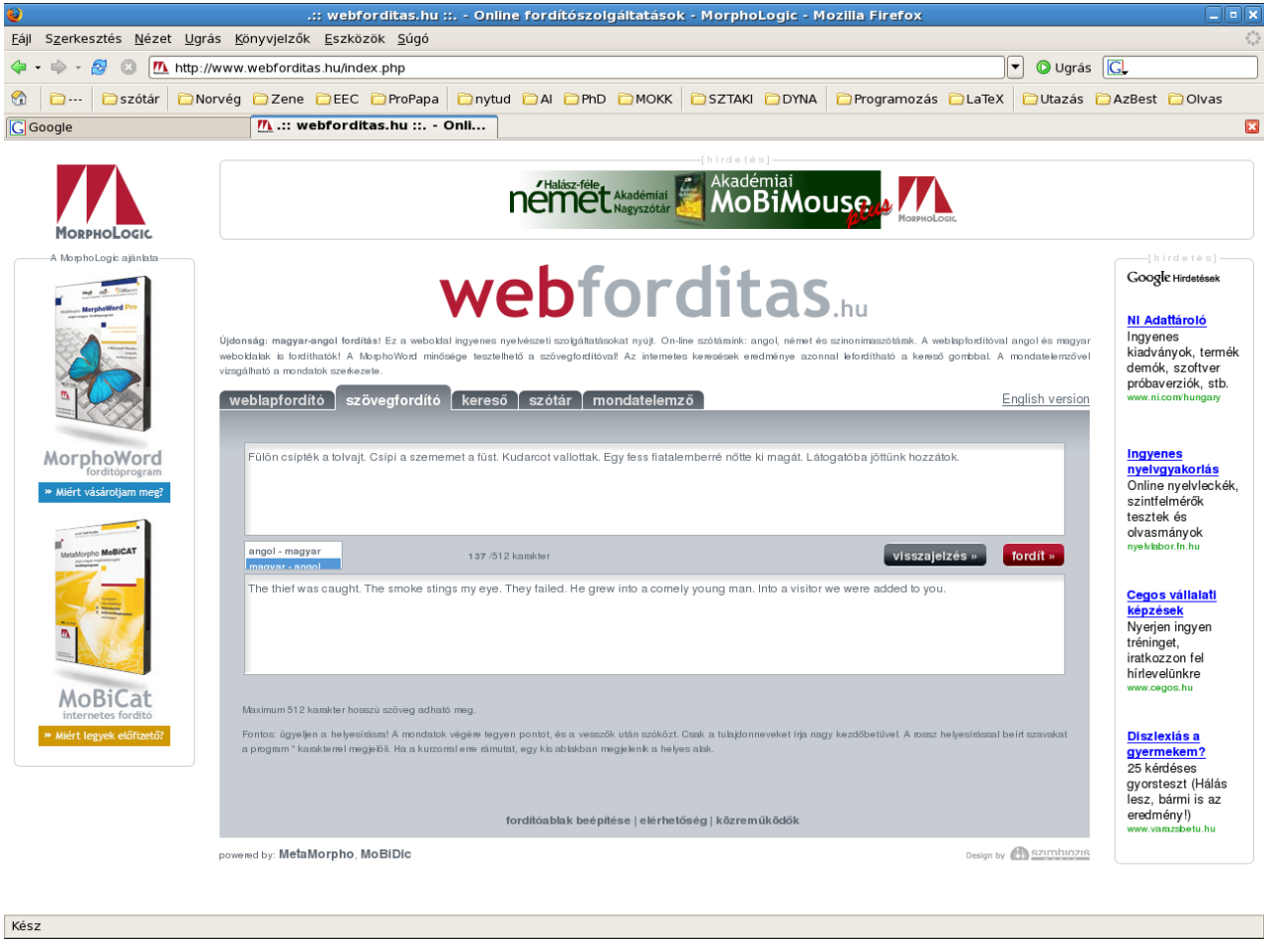
KERET-BÖNGÉSZŐ: „MAZSOLA”



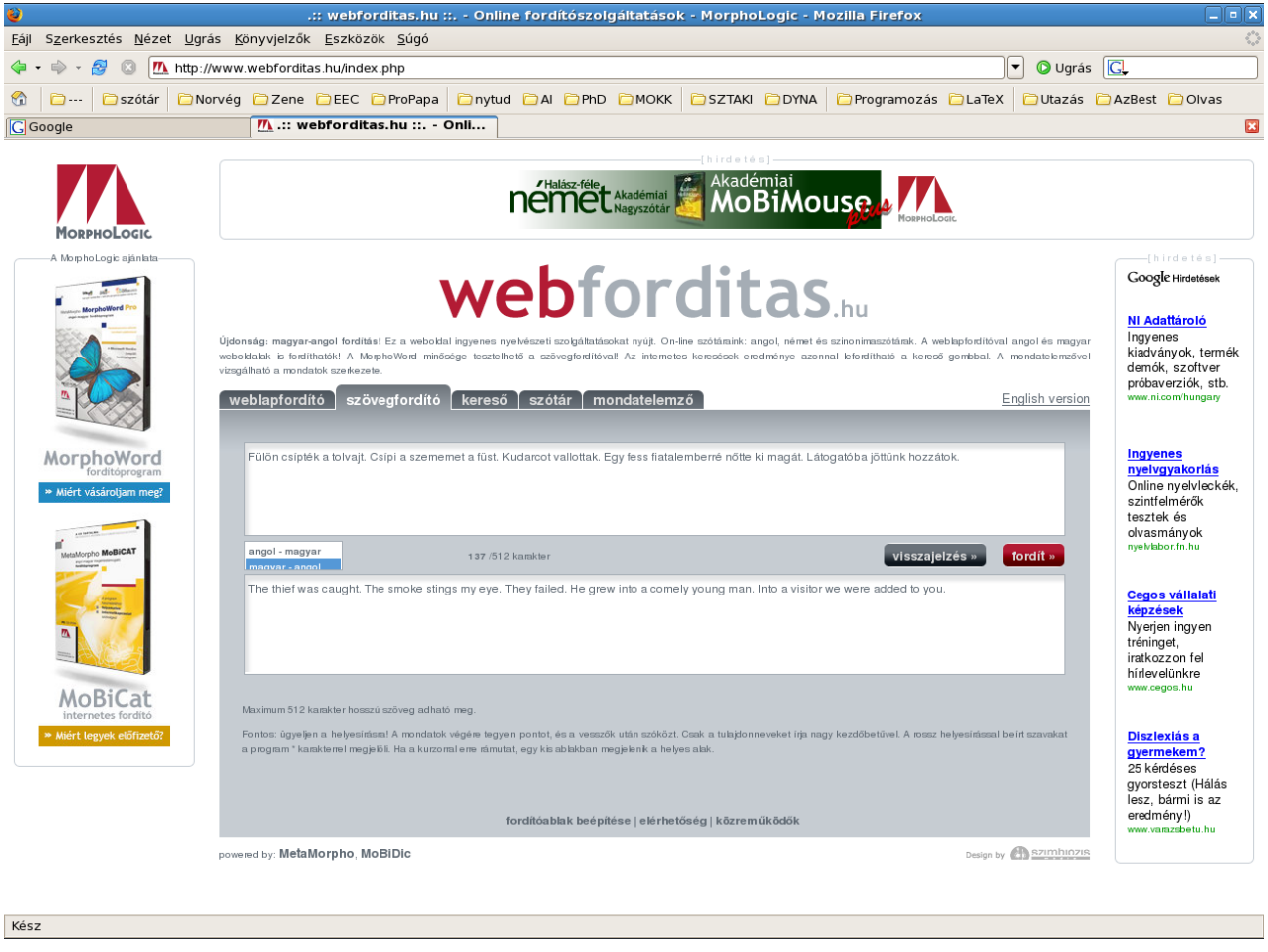
`http://corpus.nytud.hu/mazsola`

Felhasználói név: `pazmanynap` – jelszó: `2007`

ALKALMAZÁS: GÉPI FORDÍTÁS



Alkalmazás: gépi fordítás



http://www.webforditas.hu

Alkalmazás: gépi fordítás

Fülön csípték a tolvajt.

Csípi a szememet a füst.

Kudarcot vallottak.

Egy fess fiatalemberré 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!