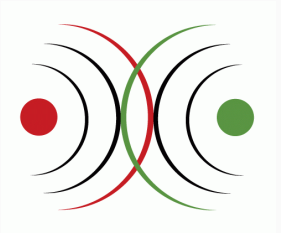


Magyar nyelvtechnológiai infrastruktúra a társadalomtudományok szolgálatában

Simon Eszter – Váradi Tamás

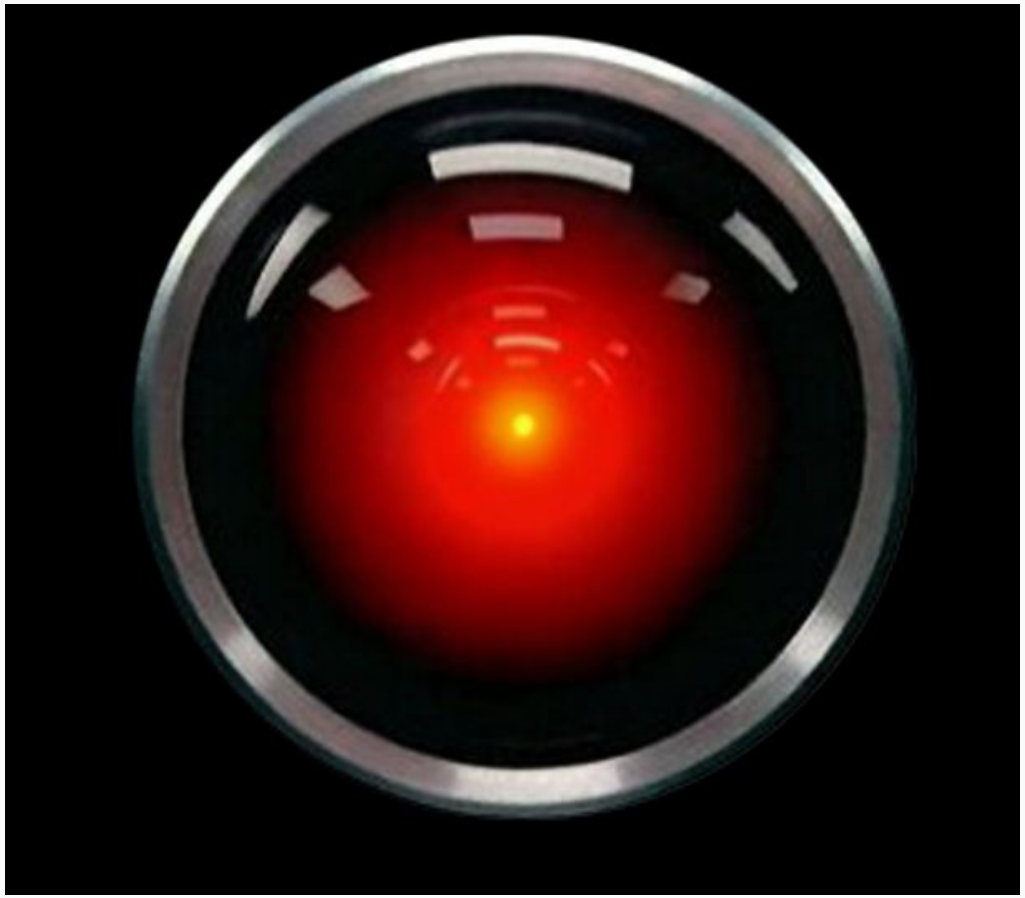
MTA Nyelvtudományi Intézet

Az előadás vázlata

1. Bevezetés
2. Nyelvfeldolgozás
3. Az `e-magyar` nyelvfeldolgozó eszközlánc
4. Alkalmazási területek
5. Összefoglalás

Bevezetés

A Szent Grál



A nyelvtechnológia célja

olyan rendszer építése, amely fel tudja dolgozni és elő tudja állítani az emberi nyelvet – úgy, ahogy az ember teszi

elméleti motiváció: az emberi nyelvhasználatot leíró formalizált és konzisztens nyelvi modellek létrehozása

gyakorlati motiváció: a modellek gyakorlati, számítógépes megvalósítása $\rightarrow$ praktikus gépi alkalmazások

Alkalmazások

a nyelvtechnológiai fejlesztések tipikusan nagyobb alkalmazásokba beépítve jelennek meg

- helyesírás-ellenőrzés szövegszerkesztőben
- természetes nyelvű keresés a böngészőben
- gépi fordítás a böngészőben
- automatikus beszédgenerálás a GPS alkalmazásban
- diktálás írott szöveggé alakítása a mobilon

Bemutatkozás I.

MTA Nyelvtudományi Intézet
Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály
Nyelvtechnológiai Kutatócsoport

- elődje: az 1997-ben alakult Korpusznyelvészeti Osztály
- osztályvezető: Váradi Tamás
- nyelvi erőforrások fejlesztése, a magyar nyelvre működő nyelvfeldolgozó módszerek és eszközök kidolgozása

Bemutatkozás II.

a Nyelv- és Beszédtechnológiai Platform és a HunCLARIN koordinátora

Nyelv- és Beszédtechnológiai Platform

- a magyarországi nyelv- és beszédtechnológiai műhelyeket tömörítő szervezet
- az akadémiai/egyetemi K+F vezető műhelyei és ipari partnerek stratégiai szövetsége
- Stratégiai kutatási terv & Megvalósítási terv
- 2008–2010, NKTH

HunCLARIN

- **CLARIN**: **C**ommon **La**nguage **R**esources and Technology **In**frastructure
- a bölcsészet- és társadalomtudományi kutatások támogatása nyelvtechnológiai eszközök és nyelvi erőforrások elérhetővé tételével
- HunCLARIN: a CLARIN magyar hálózata $\rightarrow$ nyelvtechnológiai támogatás a magyar kutatásokhoz

Nyelvfeldolgozás

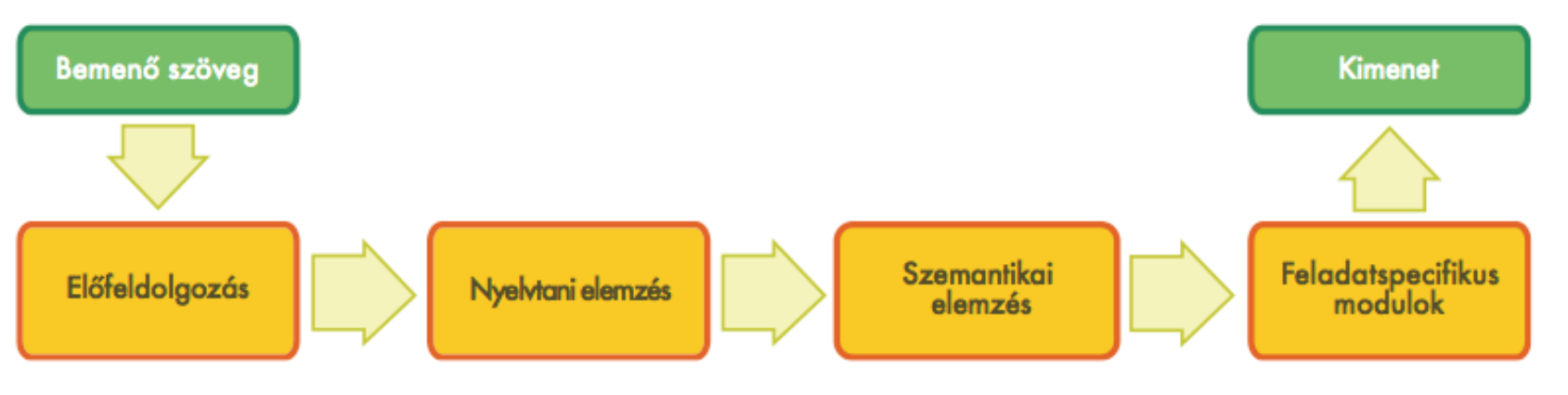
A nyelvtechnológia feladata

- bár a nyelvtechnológia folyamatosan fejlődik, még távolról sem tekinthető megoldottnak a nyelvfeldolgozás minden lépése ← az emberi nyelv komplexitása
- az egyik fő nehézség: az ember az értelmezés során számos nehezen formalizálható tényezőt is figyelembe vesz
  - a megnyilatkozás körülményei (hol, mikor, kikkel)
  - többletjelentés (ígéret, fenyegetés, irónia)
- a nyelvtechnológia feladata jelenleg: a szövegfolyamban detektálható releváns információ adott célnak megfelelő feldolgozása

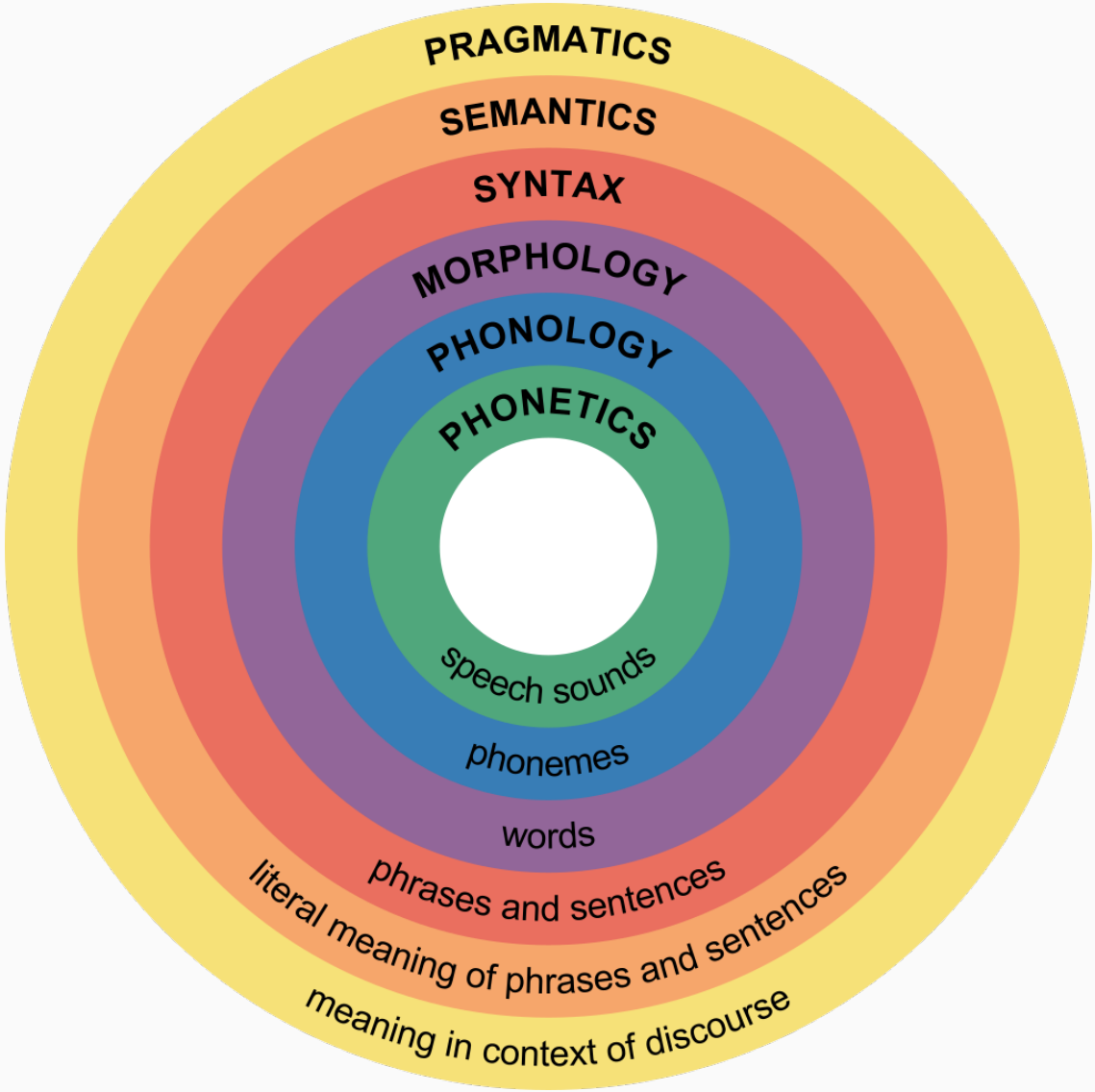
A nyelvfeldolgozás lépései

1. a digitális adatfolyam automatikus feldolgozása
2. az eredeti anyagban expliciten nem szereplő információ megtalálása
3. az adatok strukturált formába szervezése
4. az eredményeknek a felhasználó számára optimális prezentálása

Egy tipikus szövegfeldolgozó alkalmazás felépítése



A nyelvi elemzés szintjei



Részfeladatok

- mondatokra bontás
- szavakra bontás (tokenizálás)
- morfológiai elemzés
- morfológiai egyértelműsítés
- szintaktikai elemzés
- tulajdonnév-felismerés
- *koreferenciafeloldás*
- *mondatok közötti összefüggések felismerése*
- *szemantikai relációk detektálása*
- *érzelmek detekciója*

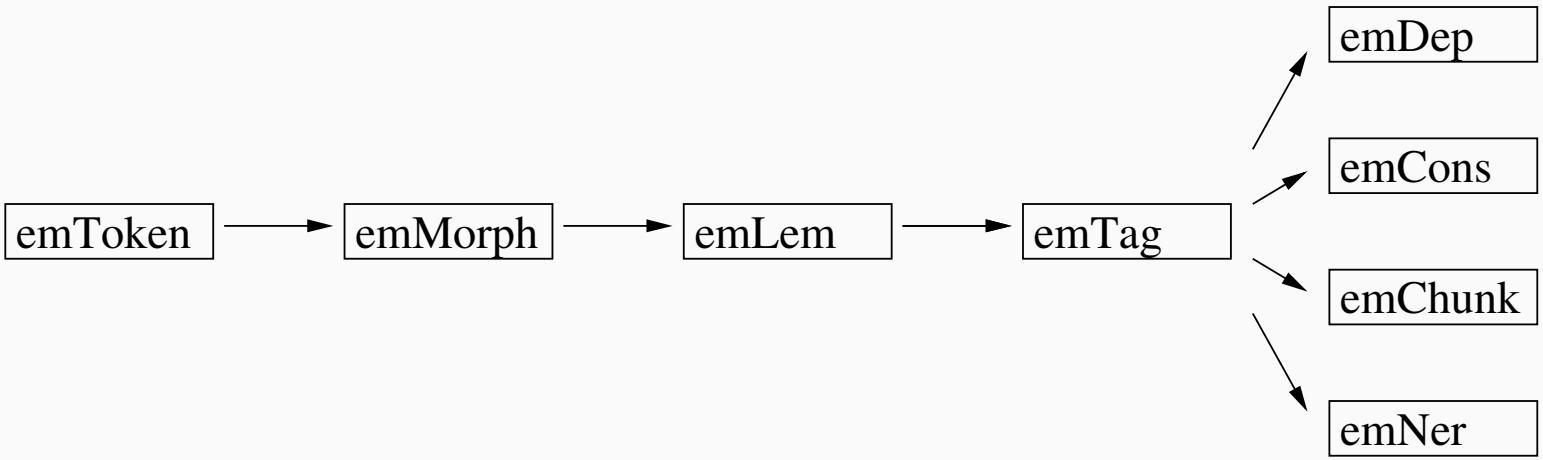
State of the art

szavakra és mondatokra bontás	$\sim 98\%$
morfológiai egyértelműsítés	$\sim 98\%$
tulajdonnév-felismerés	$\sim 95\%$
főnévi csoportok felismerése	$\sim 94\%$
vonzatkeretek kinyerése	$\sim 65\%$
metaforikus kifejezések detektálása	$\sim 43\%$

Az `e-magyar` nyelvfeldolgozó eszközlánc

Az e-magyar nyelvfeldolgozó eszközlánc

az alapvető szövegfeldolgozási lépéseket valósítja meg



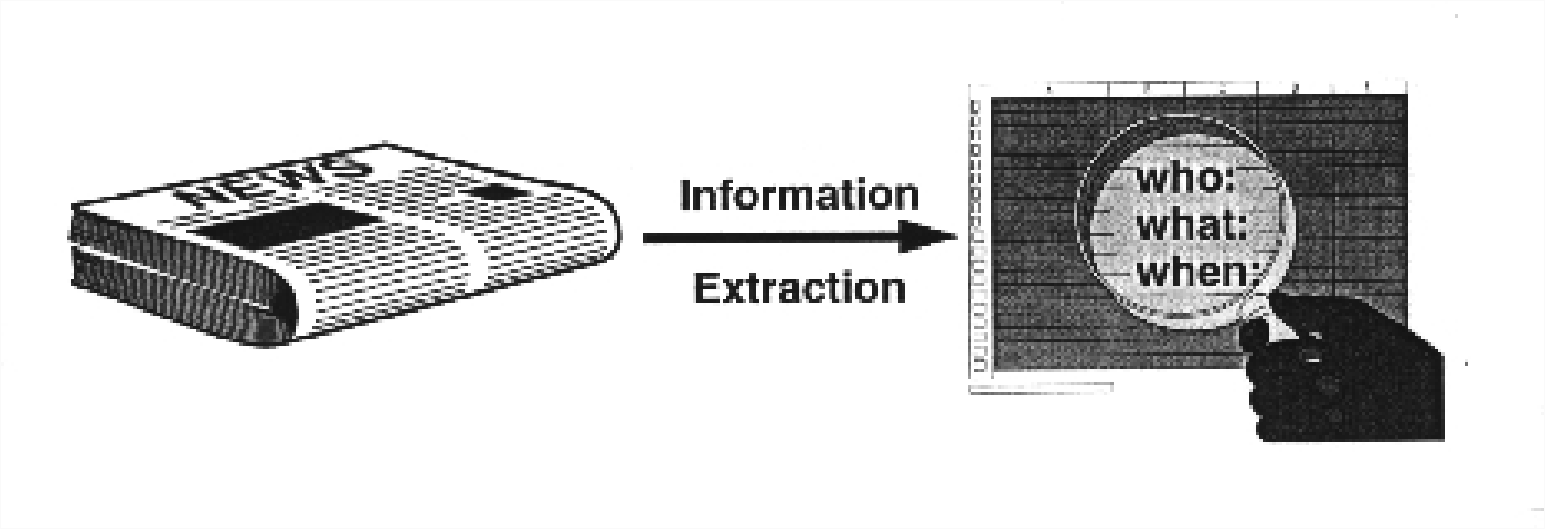
Kulcsszavak

- integráció: a különböző műhelyekben eddig előállított eszközök egy láncba szervezése
- interoperabilitás: az eszközök közötti átjárhatóságot a GATE teremti meg
- modularitás: az infrastruktúra egyes eszközei önállóan is működnek, de láncba is szervezhetők
- nyílt rendszer: az infrastruktúra egésze és minden eszköze külön-külön is elérhető
- széleskörű elérhetőség: nemcsak a nyelvtechnológusoknak, hanem kutatóknak és akár laikus érdeklődőknek is szánjuk

http://e-magyar.hu/

Alkalmazási területek

Információkinyerés



az információözönből megtalálni a releváns információt, és csak azt

az eredmény strukturált tárolása és megjelenítése

Felhasználási lehetőségek

- álláspiaci információkinyerés, trendelemzés
  - álláshirdetések és önéletrajzok automatikus begyűjtése, adatbázisba rendezése → gyorsabb és hatékonyabb egymásratalálás
- klinikai információkinyerés
  - kórházi vizsgálati dokumentumok, zárójelentések feldolgozása → hasznos információk, statisztika az egészségügy, a gyógyszeripar számára
- hangzó anyagokból történő információkinyerés
  - beszédfelismerés → szöveges átirat → szöveges információkinyerő technikák
- véleménykinyerés közösségi tartalmakból
  - blogok, fórumok, bejegyzések feldolgozása → termékekkel, pártokkal, közszereplőkkel kapcsolatos vélemények detektálása

Nyelvalapú diagnosztika

a kóros eseteket tükröző ún. nyelvi markerek keresése beszédben vagy szövegben

Felhasználási lehetőségek

- az Alzheimer-kór korai diagnosztizálása beszédtechnológiai fejlesztésekkel
  - a spontán beszéd egyes paraméterei (szünetek, agrammatikus kifejezések) a rövidtávú munkamemória teljesítményéről árulkodnak
- pszichodiagnosztikai vizsgálatok nyelvtechnológiai támogatással
  - a pszichológiai folyamatok a verbális viselkedésben is kódolódnak $\rightarrow$ tartalomelemzés $\rightarrow$ konfliktus-előrejelzés
- korpuszalapú gyereknyelvi kutatások
  - a tipikus fejlődésű gyerekek nyelvének vizsgálata segít az atipikus nyelvi fejlődésű csoportok nyelvi diagnózisában és a fejlesztés kidolgozásában is

A nyelvi kulturális tartalmak digitalizálása

Európa gazdag kulturális örökségének ápolása, megőrzése és minél szélesebb közönséggel való megismertetése

Felhasználási lehetőségek

- magyar nyelvváltozatok adatbázisa
  - Magyarországon: a dialektusok eltűnése, határon túl: nyelvvesztés
  - beszélt nyelvi anyag szöveges átirattal → fonetikai, szociolingvisztikai kutatások
- hang/film/multimédia archívumok szövegtartalom szerinti kereshetővé tétele
  - beszédfelismerési technológia → indexálás → információ-visszakeresés
- rokon nyelvek nyelvi erőforrásainak fejlesztése
  - az erősen veszélyeztetett vagy kihalt rokon nyelvek nyelvi adatainak modern eszközökkel való rögzítése, illetve a meglévő anyagok digitalizálása, elérhetővé és kereshetővé tétele
- a régi magyar nyelvemlékek digitális korba való átmentése
  - egyszerű digitalizálás: a primér adat képként való beszkennelése ↔ szöveges adatbázisok: nyelvészeti annotációval ellátva

Összefoglalás

Összefoglalás

- az alapvető szövegfeldolgozási lépések a magyarra már megoldottnak tekinthetők ↔ a magasabb szintű elemzés a nyelv komplexitása miatt nehézségekbe ütközik
- `e-magyar` nyelvfeldolgozó eszközlánc
- további alkalmazási területek, rendszerint nagyobb rendszerekbe beépítve

az MTA Nyelvtudományi Intézet koordinálásával, a HunCLARIN égisze alatt a magyar nyelvtechnológia kész és képes hatékonyan segíteni a társadalomtudományi kutatásokat

Köszönöm a figyelmet!

simon.eszter@nytud.mta.hu
varadi.tamas@nytud.mta.hu

http://e-magyar.hu