Megértéstámogatás: a nyelvi információ és környezete

Prószéky Gábor
proszeky@morphologic.hu

MorphoLogic
http://www.morphologic.hu

Pázmány Péter Katolikus Egyetem
Információs Technológiai Kar
http://www.itk.ppke.hu

Tartalom

- A megértés fogalmáról
- A környezet fogalmáról
- A jelentésről
- A szótárról
- A többszavas kifejezésekről
- A gépi szövegelemzésről
- A környezetről mint a gépi nyelvészet egyik kulcsfogalmáról
- A megszorítás-alapú rendszerekről
- … és a fordításról

A megértés fogalmáról

- Megértés = (nyelvi) információ + megfelelő „méretű” környezet
- Az életben a környezet a három dimenzióból és az időből áll
- A nyelvben a tárgyalási univerzumból
- Példa:

  Minden kutya háziállat.
  Minden kutya ugat.

  A kutya háziállat.
  A kutya ugat.
- Ha valamire azt mondja valaki: „*Nem értem*”, az azt jelenti, hogy „*Nem elég a környezet*”

A környezet fogalmáról

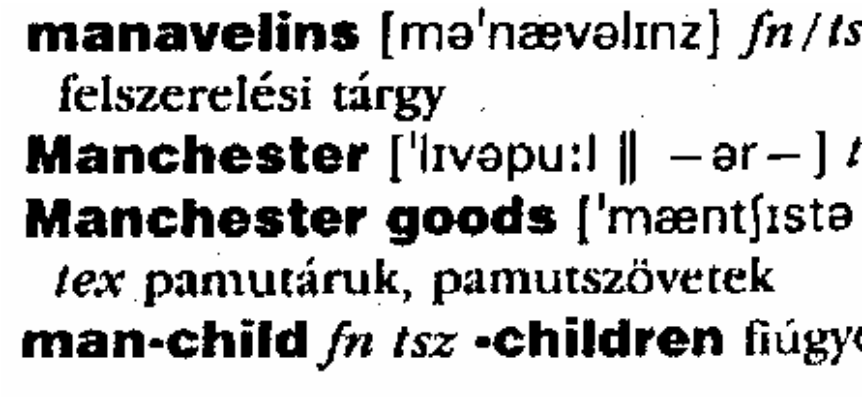
- Szintaxis = szó a környezetében
- A szó a mondatban vagy alá- vagy fölé van rendelve más szavaknak
- … azoknak a szavaknak, melyek a környezetét alkotják
- Következmény:
  nincs szó önmagában, csak a többiek között
- A szó csak a többiek között jelenti azt, amit jelent
- Más környezetben mást jelent:
  csak azért, mert más szavak vannak körülötte
- Mi van, ha nem minden ismert a környezetében?
- Nyelven kívüli-e az információ, ha ez lexikális kérdés?

A jelentésekről

- Összegyűjtjük a jelentéseket: ez (lenne) a szótár
- Jelöljük is, hogy melyik jelentés mire vonatkozik valójában?
- Nem.
- Pontosabban csak a harmadik-negyedik jelentést jelölik, de miért csak azt?
- A szótár = a szó használatának más (hasonló kultúrkörbe tartozó) emberek számára történő leírása
- Milyenek is a szótárak?

Egy szótári címszónak (és környezetének) elemzése

- Ilyenek a szótári szócikkek a nyomtatott szótárakban:


manavelins [məˈnævəlɪnz] *fn / ts*
felszerelési tárgy
Manchester [ˈlɪvəpu:l ‖ –ər–] *t*
Manchester goods [ˈmæntʃɪstə
tex pamutáruk, pamutszövetek
man-child *fn tsz* **-children** fiúgy

- Kérdés: kinek készül a szótár?
- Válasz: akinek a fejében ott a teljes szövegkörnyezet
- Ilyen eddig csak az ember volt: a kocsit is csak a ló húzhatta – egészen az autó feltalálásáig!

A szótár kifordításáról

- Szótár-kifordítás: egyirányú szótár „környezettel”
- Ez a környezet a jelentések mentén alakul, nem az ábécérend mentén
- A hasonló szavak „tere” is környezet
- A többszavas kifejezések több helyről „szedendők össze”
- Többszavas kifejezések: idiómák, kollokációk, -izmusok

A többszavas kifejezések felismerésének problémája

- Mi „szokott” mivel együtt járni?
- Együtt gyakoribb, mint külön (kölcsönös információ)
- A webkorpusz segít (korábban ilyen nem volt)
- Kell-e hozzájuk morfo-szintaktikai leírás?
- Történetileg alakultak ki: együtt állnak – és kész!
- Adekvát (szinkron) nyelvészeti leírásuk:
  a konstrukció
- A pszichológusok ezt eddig is ismerték:
  Gestalt
- Pszicholingvisztika:
  egészleges és analitikus hozzáférés
- Mesterséges intelligencia:
  a forgatókönyv magasabb szintű leírásai

Az aktuális elemzés mélységéről

- Miért akar a nyelvészet „lemenni” elemzéskor a legalsó szintre is?
- A nyelvi elemzés „granularitása” más a mondatszerkezetben, mint az etimológiában
- Példák:
  - Ki „hallja meg” ma a katonai szakkifejezéseket a viták szóhasználatában? Pl. „állást foglal”, „ütközteti a véleményeket”, „védekezés”, „lefegyverző”, …
  - Korpusznyelvészeti evidencia: a melléknévi igeneves szerkezetek sokkal kevesebb bővítményt tartalmazhatnak, mint ugyanezek az igék állítmányi szerepben

A nyelvi információ környezetei – a gépi nyelvészet szemszögéből

- A nyelvi környezet:
  konkrét, „megfogható”
- A 3D-környezet:
  konkrét, gépen kívüli
- Az időbeli környezet:
  konkrét, „megfogható, ha kell”
- A világismereti környezet:
  általános, részben modellálható
- A kulturális környezet:
  általános, még ember számára is nehezen modellálható
  (pl. japán meghívás)

Ami géppel modellálható

- A világismeret egy része
- A pragmatikai viselkedés egy része
- A nyelvi környezet viszont teljesen!
- Eddig még nem volt ilyen (!), ui. „valaki más” is ugyanazt látja, mint az adott kommunikációs helyzetben levő ember
- Ott a képernyőn az a szöveg, amit én is meg szeretnék érteni, és a gép is

A környezet felismerése más ismert környezetek alapján

- A gépi tevékenység: a keresés
- Ismét a korpusznyelvészet
- Nyelvi anyagban keresünk: nyelvi intelligencia kell
- Melyik szöveg hasonló – és miért?
- Mértékek a hasonlóságra
- Webes keresés, fordítómemóriabeli keresés

A megértés kulcsai

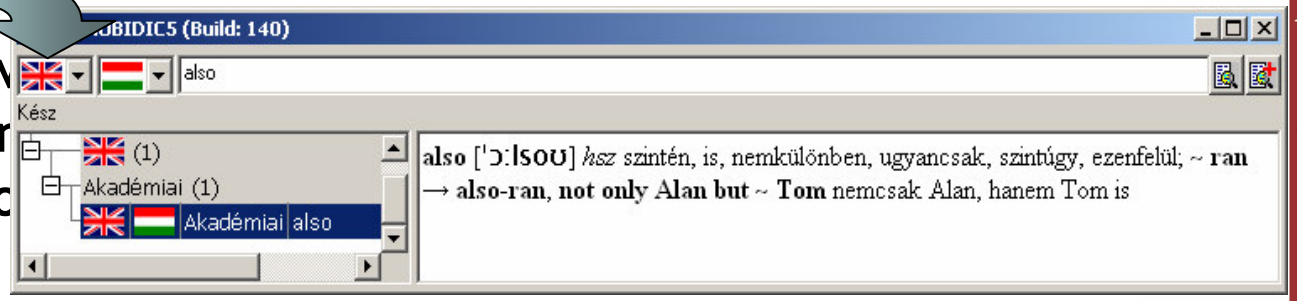
- A keresés technikai megvalósítása
- Az adott nyelv ismerete (bár olykor anyanyelvünkön is toleránsnak kell lennünk)
- A szótárban való keresés egyedi tulajdonságai
- Tudni kell, hogy amit találunk, az akkor és ott adekvát-e, „ahonnan” és „amikor” keressük (az aktuális környezetből)
- Tehát a kiinduló szöveg interakcióba lép a mi statikus(nak gondolt) tárunkkal

Intelligens szótárak: „környezettárak”

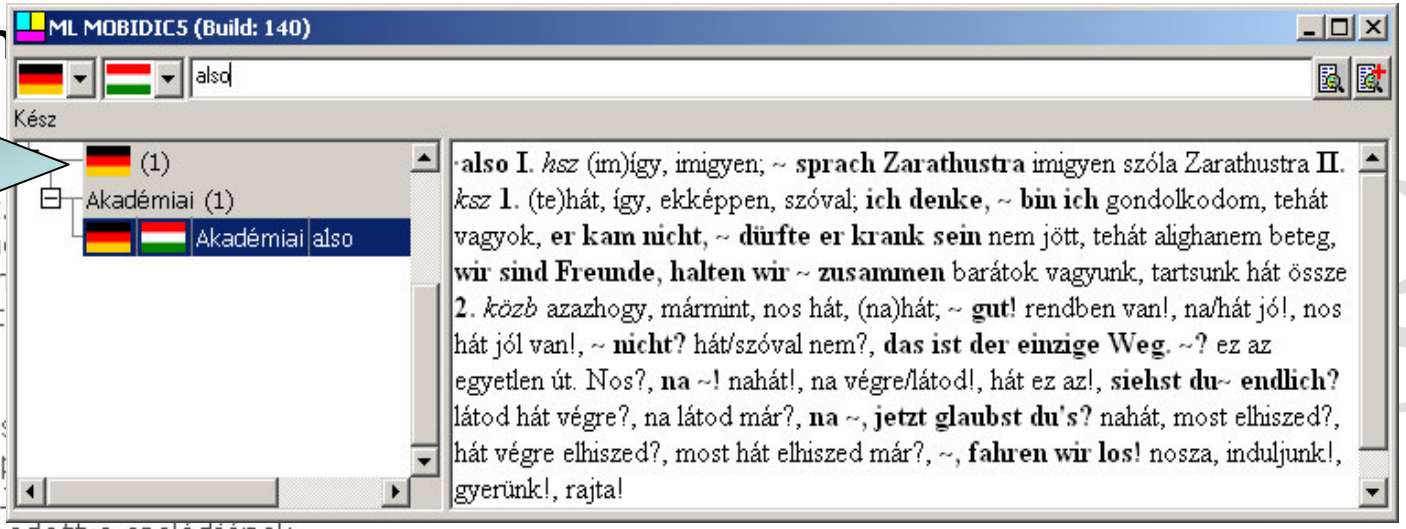
- Speciális belső reprezentációt igényelnek
- Az alszócikkek szerepe a dinamikus szócikk-építésben
- Többszörös keresés: a keresendő szót tartalmazó kifejezések más címszóban is előfordulhatnak (a szónak sokféle környezete lehet!)
- Tövesítve keresünk: egy jó szótárban ugyanis lényegesen több a ragozott szó, mint a ragozatlan (!)

Példák a környezetérzékeny szótár viselkedésére

By ‘signifi t w
there are also ‘r
skipped when fo

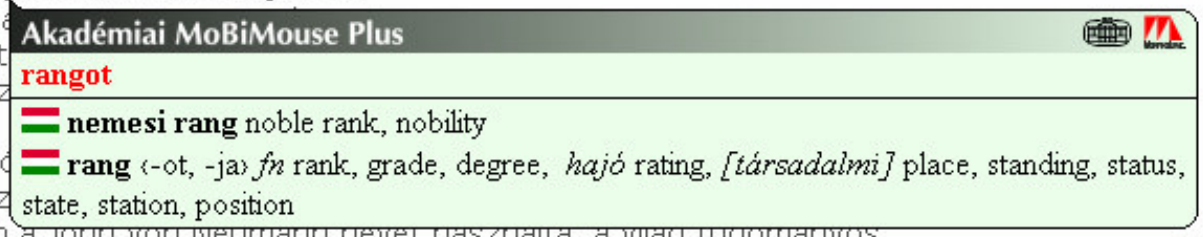


Also sprach



születésnapját.
alkalom arra, h
körben megism
tudós életútját
élő hatását.

Neumann János
született Budap
amikor Ferenc J
nemesi rangot adott a családjának.
Apja, dr. Ma
anyagi jólét
alapokat biz
elsőszülött
fiúnak, Miklós
nyugszik az
Amerikában a John von Neumann nevet használta, a világ tudományos
élete ezen a néven ismeri.



Mondatjelentés: szavak környezeteikkel

- Kifejezések fordítása minden mondatbeli szóra
- Egy k hosszú mondatban mind a k szónak a többi k-1 szó a környezete
- A szavak nem feltétlenül literálisan értendők, olykor jelentésbesorolásuk, szófajuk is elegendő
- Szedjük össze ezeket a részfordításokat, és „rakjuk őket egymásra”

Megszorítások: a környezet használata

- Az informatikában jó ideje jelen van a „constraint” alapú gondolkodás (pl. szakértő rendszerek)
- A nyelvoktatásban:
  - „ez és ez a szerkezet azt jelenti, hogy X, mert az angol (francia, …) így mondja
  – meg kell tanulni az anglicizmusokat (x-izmusokat), kifejezéseket, idiómákat!”
- Most pedig a szövegek megértésének modellálásában jelenik meg a megszorítás

Megszorítások használata a gépi nyelvészetben

- Karlsson „Constraint Grammar”-ja
- A konstrukciós nyelvtanok
- Koskenniemi kétszintes morfológiája
  - Környezetfüggőség és megszorítás:
    (a) leíró nyelvészeti szabadság (CS-szabályok) és
    (b) hatékony számítógépes megvalósítás
    (reguláris kifejezések)
- Kulcs: a lokális környezet és az általános környezet különbségének felismerése

Mi tehát a fordítás?

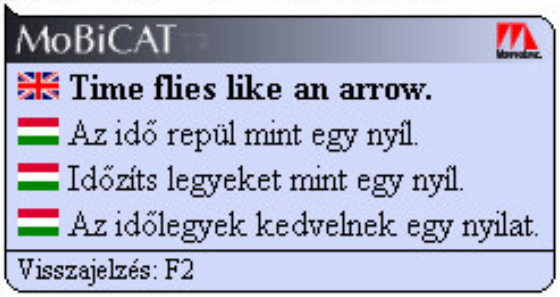
- A mondat jelentése egy másik nyelv szavaival kifejezve
- Fordítás = megértés + ennek a kifejezése egy másik formalizmusban
- A számítógéppel való fordítást vissza lehet tehát vezetni E. Bach „rule-to-rule” hipotézisére
- Ezt eddig csak a logikai szemantika szintaktikai szabályokhoz rendelésénél használták
- Minden szerkezetnek megvan a maga fordítása: ez a (kissé általánosított) szótár
- A mondat fordítása a részszerkezetek fordításainak egy függvénye
- A kifejezés jelentése pedig tényleg annak használati szabálya

Példa: a megszorításos gépi fordítás menete

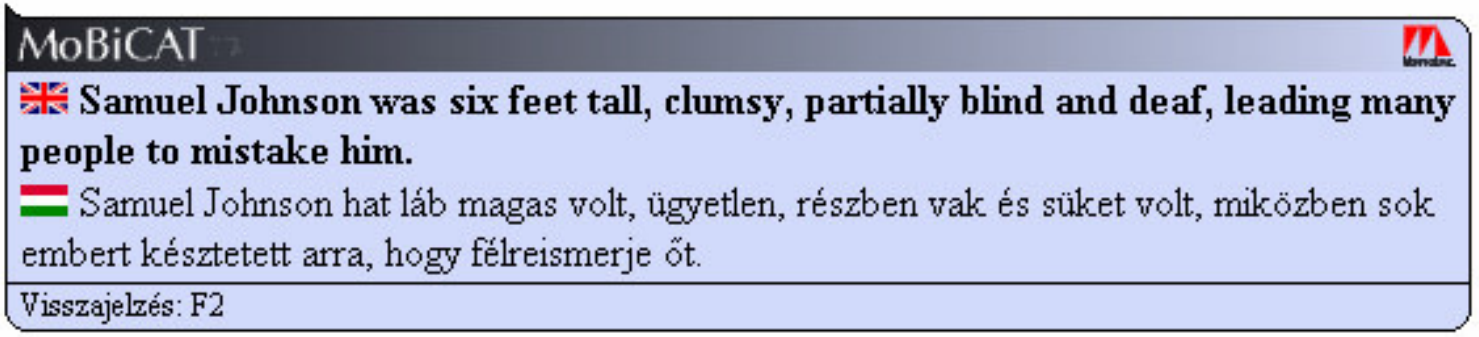
```
1---2----3---4----5-----6--7-------8
Jim does not sink money in anything.
1--
Jim
    2----
    tesz
         3---
         nem
             4------
             süllyed
             süllyeszt
    2----3---4----
    nem süllyed
    nem süllyeszt
                  5-----
                  pénz
                           7--------
                           bármi
             4----5-----6--7--------
             befektet pénzt bármibe
             4----5----
             pénzt süllyeszt el
             befektet pénzt
    2----3---4----5-----
    nem fektet be pénzt
    2----3---4----5-----6--7------
    nem fektet be pénzt semmibe
1---2----3---4----5-----6--7------8
Jim nem fektet be pénzt semmibe.
```

Példa: a megszorításos gépi fordítás a gyakorlatban

Time flies like an arrow.

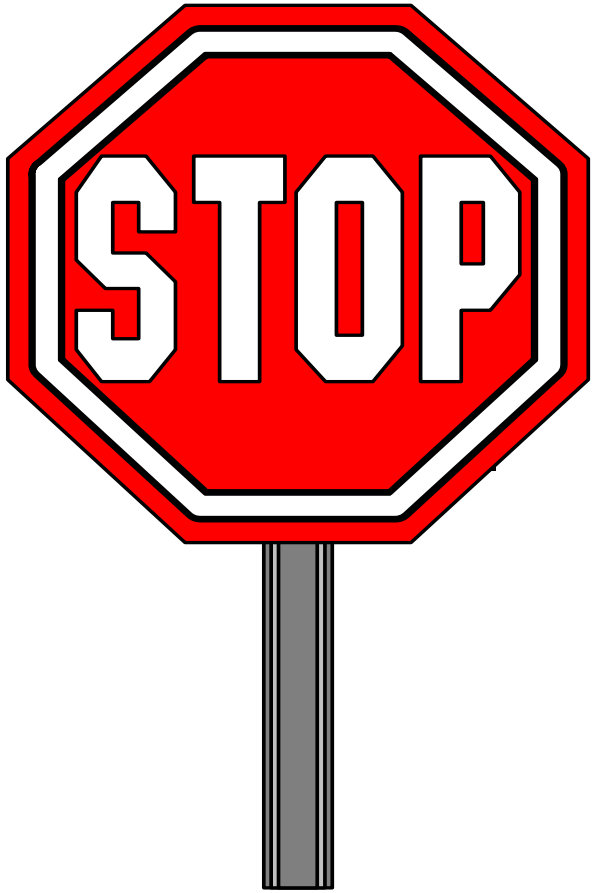


Samuel Johnson was six feet tall, clumsy, partially blind and deaf, leading many people to mistake him.



A(z informálisan) javasolt/megvalósított modellek összefoglalása

- Szótármodell
  ami dinamikus, azaz az aktuális környezetre érzékeny
- Fordításmodell
  ahol a fordítás a forrásnyelvi rész-szerkezetek fordításainak függvénye
- Nyelvmodell
  amelyben az ábécé nyílt halmaz, és megoldja az ismeretlen szavak kezelésének problematikáját


STOP