

A 4lang fogalmi szótár

Kornai András és Makrai Márton

MTA SZTAKI Nyelvtechnológiai Kutatócsoport

MSZNY 2013.

Áttekintés

Bevezetés

A definíciók szintaxisa

A definiáló szókincs jellemzése

Más lexikai adatbázisokkal összehasonlítva

Más lexikai adatbázisokkal összehasonlítva

- ▶ szavak absztrakt jelentését formalizálja (monoszémia)
 - ▶ egyértelműsítés lehetőleg csak tiszta homonímia esetén (state₇₆ 'állam', state₇₇ 'állapot')
 - ▶ szófajmentes szemantika
- ▶ fogalmak közötti kapcsolatokat rögzít, nem a világról való ismereteket
- ▶ többnyelvű, lehetőleg nyelvfüggetlen

A tételek felépítése, több nyelv

A tételek felépítése, több nyelv

| | | | | | | |
|-----|---------|-----|-------|---------|--------------|------------------|
| 102 | átenged | V | pass | concedo | przepuścić : | LET[DAT HAS ACC] |
| id | magyar | POS | angol | latin | lengyel | definíció |

A tételek felépítése, több nyelv

| | | | | | | |
|-----|---------|-----|-------|---------|--------------|------------------|
| 102 | átenged | V | pass | concedo | przepuścić : | LET[DAT HAS ACC] |
| id | magyar | POS | angol | latin | lengyel | definíció |

- ▶ 40 nyelvre való kiterjesztés folyamatban van

Áttekintés

Bevezetés

A definíciók szintaxisa

A definiáló szókincs jellemzése

- ▶ egyváltozós predikátumok
 - 1474 lány N girl puella dziewczyna: female, child
 - 112 acél N steel chalybs stal: metal, hard, strong
- ▶ kétváltozós predikátumok
 - 1656 mell N breast mamma pierś: two, organ, breast ON chest, woman HAS breast
 - 1233 kard N sword gladius miecz: weapon, sword HAS blade[<long>,pointed], sword HAS edge
- ▶ mélyesetek
 - 102 átenged V pass concedo przepuścić: NOM LET[DAT HAS ACC]
 - 2374 tesz V put pono kłaść: NOM CAUSE[ACC AT OBL], NOM MOVE ACC, ACC[object]
- ▶ többargumentumúak visszavezetése kétargumentumúakra

- ▶ alapértelmezett (*default*)
1614 medence N pool piscina basen: water IN, <swim>
IN, <play> IN
1724 mos V wash lavo mycí: CAUSE[ACC[*clean*]],
INSTRUMENT liquid, INSTRUMENT <soap>, INSTRUMENT rub
- ▶ „eseményszerkezet”: before[], after[]
715 fagy N freeze gelu mróz: cold CAUSE,
before[liquid], after[solid,<ice>]
2616 vezet V guide rego prowadzić: CAUSE[ACC HAS
information], information ABOUT place₁₀₂₆, after[ACC AT
place₁₀₂₆]
- ▶ tagadás – az alapértelmezettől való eltérés
500 e1j N night nox noc: period, FOLLOW sunset,
sunrise FOLLOW, dark, lack(sun), <sleep AT>
931 gyerek N child puer/puella dziecko: person,
young, lack(responsible), parent MAKE

Áttekintés

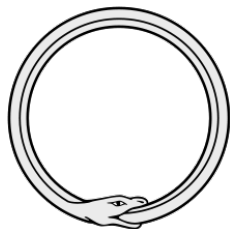
Bevezetés

A definíciók szintaxisa

A definiáló szókincs jellemzése

Az alapszókincs

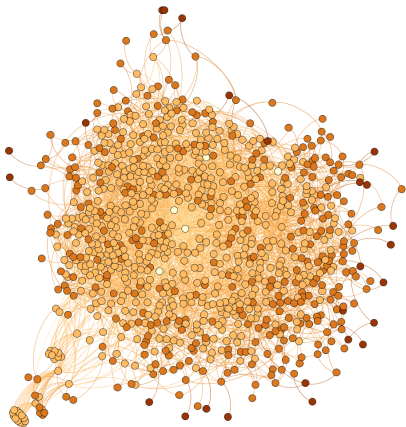
- ▶ a lexikográfia régi problémája a körkörösség, *valódi* \Leftrightarrow *létezik*
- ▶ a szótáraknál szokásos megoldás: definiáló szókincs (DV)
- ▶ Longman: 82 128 \rightsquigarrow 2 960
- ▶ 4lang: 2 960 \rightsquigarrow 1 156
- ▶ uroborosz tulajdonságú szóhalmaz
- ▶ maradnak primitívek
- ▶ maradnak körök
- ▶ hivatkozunk érzékletekre (*függőleges*) és az enciklopédia elemeire (*baseball*)
- ▶ kötött morfémák (*seventh*)



A definíciós gráf



A definíciós gráf



- ▶ csúcsok: fogalmak, 2 897 db
- ▶ irányított élek: 'acél' → 'fém', 7816 db

- ▶ a szókincs súlyozása a definíciókban való fontosság szerint
- ▶ véletlen séta határeloszlása
- ▶ a véletlen séta határeloszlása egyértelmű \Leftrightarrow a gráf erősen összefüggő
- ▶ egy u és egy v csúcs *erősen összefüggő*, ha van $u \rightsquigarrow v$ és $v \rightsquigarrow u$ út
- ▶ ekvivalenciareláció, komponensei az *erősen összefüggő komponensek*

A 4lang gráf erősen összefüggő komponensei

| méret | db | |
|-------|------|---|
| 662 | 1 | {yellow, four, sleep, under, lack, month...} |
| 12 | 1 | {január, február, ..., december} |
| 7 | 1 | {hétfő, kedd, ..., vasárnap} |
| 5 | 1 | {furniture, chair, table, bed, cupboard} |
| 4 | 3 | {queen, royal, monarch, king}, {cereal, flour,...}... |
| 3 | 8 | {male, sex, female}, {calm, disturb, upset},... |
| 2 | 26 | {exist, real}, {reason, cause}, {child, parent},... |
| 1 | 2302 | {PART_OF}, {other}, {IS_A}, {number}, ... |

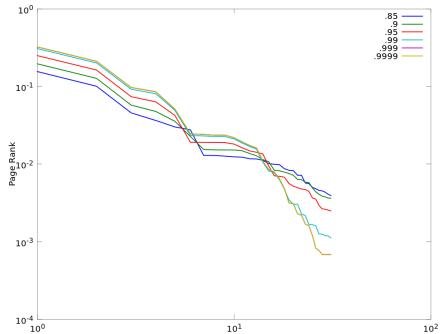
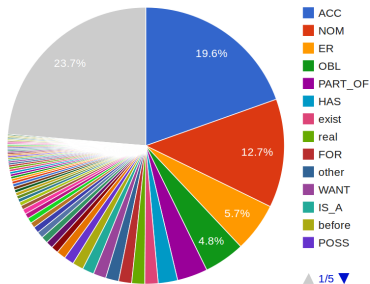
Page Rank, csillapítás

- ▶ Page Rank: az átmenetmátrixot *csillapítással (damping)* erősen összefüggővé tesszük

$$M_d(i,j) = \frac{1-d}{N} + dM(i,j)$$

ahol az M mátrix $N \times N$ -es

- ▶ $d \approx 1$ esetén a Page Rank közelíti az eredeti séta határeloszlását



- ▶ terv: a Longman szótár 82 000 szavának lefordítása a mi formalizmusunkba gépi tanulással
- ▶ <http://hlt.sztaki.hu/resources/4lang/>
- ▶ Laptopos bemutató: Miből lesz a robot MÁV-pénztáros? – Nemeskey Dávid, Recski Gábor, Zséder Attila
- ▶ Köszönöm a figyelmet!

A fogalmak Page Rankje hatványeloszlást követ?

- ▶ $p(x) \propto x^{-\alpha}$ fennáll?
- ▶ Clauset és tsai 2009
- ▶ $\alpha = 1.9244$
- ▶ $x_{min} = 2.4219 \cdot 10^{-4}$
- ▶ Kolmogorov–Smirnov-statisztika: $0.5840 > 0.1$, nem rossz
- ▶ kellne még: likelihood-arány tesztek *likelihood-ratio tests*

Definíciók bonyolultsága

- ▶ az átmenetmátrix legnagyobb jobboldali sajátértékéhez tartozó sajátvektor koordinátái

| | |
|--------------------|------------------------|
| 0.22008937659358 | mind |
| 0.1483881645837043 | read |
| 0.1419981763311443 | autumn |
| 0.1343248456669039 | brain |
| 0.1340132834187455 | feel |
| 0.1296319053350392 | understand |
| 0.1256292901034924 | remember |
| 0.1210243324308604 | summer |
| 0.1199256807742143 | sensible |
| 0.1130671228119619 | spring ₂₃₁₈ |