Fogalmak fontossága a definíciós gráf vizsgálatával

Makrai Márton

ELTE BTK Nyelvtudományi Doktori Iskola
makrai@budling.hu

Kivonat: A cikkben egy számítógépes szövegértéshez szolgáló erőforrás, a 4lang fogalmi szótár segítségével vizsgálom, hogy az egyes fogalmak mennyire fontosak a mondatmegértésben. Ezt úgy teszem, hogy a szavak jelentését reprezentáló definíciókat egy irányított gráffá alakítom, melynek csúcsai a fogalmak. A cikkben leírt definíciós gráfra a számítástudományban Page Rank néven ismert módszert alkalmazva az egyes csúcsokhoz rendelt értékek éppen a megfelelő fogalomnak a más szavak és frázisok megértésben való fontosságaként értelmezhetők.

1 Bevezetés

A számítógépes szövegértés egyre inkább átszövi a mindennapokat. Az okostelefonok természetes nyelvű kérdésekre válaszolnak, az ingyenes statisztikai alapú gépi fordítók segítségével megtudhatjuk, hogy egy általunk egyáltalán nem beszélt nyelven írt szöveg nagyjából miről szól, és a technika mostanában kezdi kielégíteni azt az igényt, hogy a gépeknek ugyanazokkal a szavakkal adhassunk feladatot, mint egy személyi titkárnak. Többek közt ebben a kutatási irányban dolgozik az MTA Számítástudományi és Automatizálási Kutatóintézetének Nyelvtechnológiai Kutatócsoportja. A cikkben a kutatócsoport szemantikaprojektjének lexikonát, a 4lang fogalmi szótárat vizsgálom abból a szempontból, hogy az egyes fogalmak mennyire fontosak a megértésben. A Page Rank módszerét használom, amit eredetileg honlapok relevanciájának mérésére találtak ki.

A cikk felépítése a következő. A második szakaszban bemutatom a 4lang fogalmi szótárat, kitérve a szókincs összeállítására és a definíciók megírásának alapelveire. A harmadik szakaszban felvázolom a szavak jelentését megragadó definíciókban használt elemeket és szintaxist. A negyedik szakaszban bemutatom a definíciós gráfot, az ötödikben pedig az egyes fogalmak súlyának kiszámítására használt módszert, a Page Ranket. Végül az ötödik szakaszban közlöm a számszerű eredményeket.

2 A 4lang fogalmi szótár

A kutatáshoz a 4lang fogalmi szótár szolgál adatként. Ez egy gépi szövegértéshez készült erőforrás, olvasója tehát egy másik számítógépes modul. A szótárat részletesen bemutattuk a IX. Magyar Számítógépes Nyelvészeti Konferencián (Kornai–Makrai 2012), így itt csak átismétlem a szókincs összeállítását és a definíciók elkészítésének legfontosabb alapelveit. További információt a fenti konferenciakötetben talál az ol-

Alknyelvdok7
Szerk.: Váradi Tamás
MTA Nyelvtudományi Intézet, Budapest, 2013
ISBN 978-963-9074-59-0

vasó. Azt, hogy az itt leírt alapelvek megvalósíthatók-e, természetesen a szótárt használó rendszer intelligenciája fogja mutatni. A hagyományos lexikográfiai előzményekkel és a számítógépes nyelvészeti ontológiákkal való kapcsolatról, valamint a kutatócsoport szemantika projektjének elméleti hátteréről lásd bővebben Kornai (2010)-et.

A 4lang gerincét (1942 darab definiált szót és kötött morfémát) a Longman Defining Vocabulary adja. A Longman szótárnak az a változata, ami rendelkezésünkre állt (Bullon 2003) ezen kívüli elemeket is használ,, ezért tovább bővítettük a szókincset 197 egyszerű szóval (pl. *dimension, two, communicate, conform, mammal, item, artefact*), 188 tulajdonnévvel, melyek definíciója lényegében csak egy hivatkozás az enciklopédia megfelelő elemére (pl. *Greenland, Greenwich, Guy Fawkes*) és 147 összetett szóval (*bell-shaped, bitter-tasting, blue-black*). Utóbbiak az itteni eredmények szempontjából érdektelenek.

A 4lang definíciói emberi munkával készültek, többnyire felhasználva klasszikus szótárakat is, leginkább a Longman szótárat. A definíciók megírásának legnagyobb újdonsága alighanem a radikális monoszémia. Ez azt jelenti, hogy a szótár a szavak *absztrakt jelentését* igyekszik megragadni, amiből a konkrét használatok levezethetők. Kornai–Makrai (2012)-ban idéztük a *Webster's Third* (Merriam 1961) *potash* 'hamuzsír, kálisó' tételét annak bemutatására, hogy a hagyományos lexikológiában hogyan szoktak egyértelműsíteni poliszémnek tartott szavakat. Náluk a *potash* szónak négy jelentése van. Az Alkalmazott Nyelvészeti Doktorandusz Konferencián egy számítógépes erőforrásból, az angol WordNet-ből (Miller 1995) hoztam hasonló példát a *stomach* 'has' szó hat jelentésével.

Noun

- S: (n) **stomach**, tummy, tum, breadbasket (an enlarged and muscular saclike organ of the alimentary canal; the principal organ of digestion)
- S: (n) abdomen, venter, **stomach**, belly (the region of the body of a vertebrate between the thorax and the pelvis)
- S: (n) **stomach** (an inclination or liking for things involving conflict or difficulty or unpleasantness) *"he had no stomach for a fight"*
- S: (n) **stomach** (an appetite for food) *"exercise gave him a good stomach for dinner"*

Verb

2 S: (v) **stomach** (bear to eat) *"He cannot stomach raw fish"*
3 S: (v) digest, endure, stick out, **stomach**, bear, stand, tolerate, support, brook, abide, suffer, put up (put up with something or somebody unpleasant) *"I cannot bear his constant criticism"; "The new secretary had to endure a lot of unprofessional remarks"; "he learned to tolerate the heat"; "She stuck out two years in a miserable marriage"*

A 4lang alapelvei szerint mindkét szó monoszém (egyjelentésű). Egyértelműsítés csak tiszta homonímia esetén történik, pl. az angol *state* szóalaknak külön tétel felel meg az 'állam' és az 'állapot' jelentésben. Kezdeti munkamódszernek azt választjuk, hogy egy szóalak egyes használatait mindaddig ugyanahhoz a jelentéshez soroljuk, amíg nem ütközünk egy olyan nyelvbe, amely a két esetben más szót alkalmaz. A momoszémiához tartozik, hogy mivel a 4lang a fogalmi jelentést írja le, egy fogalom-

hoz sorol két olyan szót, ami csak szófajában különbözik, pl. az igéket a belőlük képzett *nomen actionis*szal (*lát, látás*).

Az absztrakció után azt emelem ki, hogy a 4lang szótár nyelvfüggetlen törekszik lenni. A szavak definiálásakor igyekeztünk több nyelvet is figyelembe venni, és a fogalmak szóalakját magyarul, angolul, latinul és lengyelül is feltüntettük. A kutatócsoportban folyamatban van a szótár negyven nyelvre való kibővítése.

Végül megjegyezzük, hogy a definíciók elkészítésekor arra törekedtünk, hogy csak nyelvi tudást, úgymond analitikus igazságokat rögzítsünk. Az 'éj' 4langbeli definíciójában (lásd a következő szakaszban) szerepel, hogy nincsen nap. Ezzel azt is mondjuk, hogy ha éjszaka (mondjuk éjfélkor) feljönne a nap, azt már nem is neveznénk éjszakának.

3 A definíciók szintaxisa

A későbbiek megértéséhez szükséges, hogy szóljunk egy pár szót a definíciók szintaxisáról. Először a következő három példán keresztül bemutatom a 4lang tételeinek felépítését.

```
1474 lány N girl puella dziewczyna: female, child

112 acél N steel chalybs stal: metal, hard, strong

500 éj N night nox noc: period, FOLLOW sunset, sunrise FOLLOW, dark,
lack(sun), <sleep AT>
```

Látható, hogy a tételek azonosítóval (sorszámmal) kezdődnek, majd a magyar alak, egy szófaji címke (a korábban elmondottak miatt ez csak kisegítő információ), az angol, a latin és a lengyel alak, végül pedig, kettősponttal elválasztva maga a definíció szerepel.

A `kisbetűs` predikátumok magukért beszélnek, az első példában azt látjuk, hogy aki lány, az nő és gyerek.

A következő két definícióban kétváltozós predikátumok is szerepelnek. Ezeket csupa `NAGYBETŰVEL`, infix jelöléssel (az argumentumok közé) írjuk. A \redu{szürkével} írt szavak nem részei az erőforrásnak, hanem egy redundanciaszabály gondoskodik arról, hogy a ki nem töltött argumentumok szerepét a definiálandó szó (az első példában tehát a 'mell') töltse ki.

```
1656 mell N breast mamma pier\'s: two, organ, breast ON chest, woman
HAS breast

1233 kard N sword gladius miecz: weapon, sword HAS
blade[<long>,pointed], sword HAS edge
```

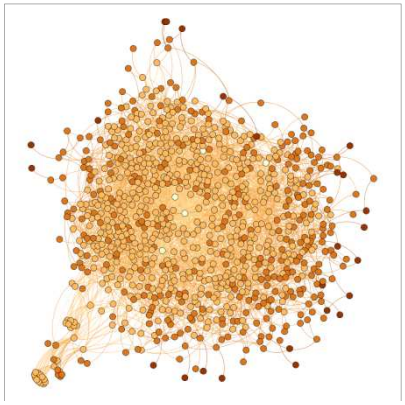
Ebben a cikkben szószinten dolgozunk, mégis fontos beszélni a szavak (tipikusan az igék és relációs főnevek) vonzatairól. A kompozícionalitás elvének speciális eseteként megköveteljük, hogy egy régensből és annak vonzataiból álló szerkezet jelentésének reprezentációja a régens és a vonzatok jelentésreprezentációjából álljon össze. Ehhez a régensek definíciójában jelölni kell, hogy hova kerül a vegyes vonzatok reprezentációja. Ezt a vonzatok mélyesetére (Fillmore 1968) hivatkozva tesszük meg. A mélyesetek elnevezése jelenleg magyar felszíni eseteket rövidít, így pl. `NOM`} jelöli az alanyt, `ACC` a tárgyat, `DAT` a datívuszi vonzatot és `OBL` az oblikvuszit. Kutatócsoportunkban folyamatban van egy elméletileg megalapozottabb (a Fillmore-i elképzeléssel

is könnyebben összevethető és alternációkat is megragadó) mélyesetrendszer kidolgozása. A példákban az is látszik, hogy a szintaxisban ditranzitív (esetleg még több argumentumú) igék jelentéséről hogyan adunk számot legfeljebb kétváltozós predikátumok segítségével. Erről részletesen ír Kornai (2012).

4 A definíciós gráf

Ebben a szakaszban először bemutatom, hogy hogyan alakítottuk át a 4lang szótárat egy irányított gráffá, melynek segítségével jellemezni tudtuk a fogalmak szemantikai fontosságát, utána pedig bemutatom a gráfot.

A definíciós gráf csúcsai a szótárban szereplő fogalmak, és ha pl. az 'acél' szó definíciójában szerepel a 'fém' szó, akkor a gráfban egy irányított él mutat az előbbinek megfelelő csúcsból az utóbbinak megfelelőbe. A gráfnak 2 897 csúcsa és 7816 éle van (tehát viszonylag ritkán vannak élek két csúcs között).



1. ábra A definíciós gráf

A későbbiekben fontos lesz az erős összefüggőség fogalma. Két csúcsot *erősen összefüggő*nek hívunk, ha vezet belőlük út (élsorozat) egymásba. Ez a reláció ekvivalenciareláció, a csúcsokat osztályokba sorolja, ezek az *erősen összefüggő komponens*ek. A 4lang gráfjának erősen összefüggő komponensei önmagukban is érdekesek, és intuíciót adnak a gráfról, ezért ezeket röviden bemutatom.

A későbbiekben fontos lesz az erős összefüggőség fogalma. Két csúcsot *erősen összefüggő*nek hívunk, ha vezet belőlük út (élsorozat) egymásba. Ez a reláció ekvivalenciareláció, a csúcsokat osztályokba sorolja, ezek az *erősen összefüggő komponens*ek. A 4lang gráfjának erősen összefüggő komponensei önmagukban is érdekesek, és intuíciót adnak a gráfról, ezért ezeket röviden bemutatom.

Ahogy a táblázatban is látszik, a legnagyobb erősen összefüggő komponensben meglehetősen vegyesen találunk szavakat. A következő legnagyobb erősen összefüggő komponenseket olyan ciklusok adják, mint a hónapok vagy a hét napjai. Ezeket úgy definiáljuk, hogy pl. minden hónapnál feltüntetjük azt, hogy ő egy hónap, valamint az előző és a következő hónapot. A következő legnagyobb erősen összefüggő komponensek valamilyen fogalomhoz kapcsolódnak értelmesen, például a bútorok és

maga a *bútor* szó egy erősen összefüggő komponenst alkotnak, ugyanis a bútorok definíciójában szerepel a bútor szó, a 'bútor'-ében pedig néhány bútor példaként. Végül a legtöbb fogalom egyedül alkot erősen összefüggő komponenst.

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, our,. . . },. . .
3	8	{male, sex, female}, {calm, disturb, upset},. . .
2	26	{exist, real}, {reason, cause}, {child, parent},. . .
1	2302	{PART_OF}, {other}, {IS_A}, {number}, . . .

1. táblázat. A 4lang erősen összefüggő komponensei

5 A fogalmak súlya

A fogalmak egymás definiálásában játszott fontosságát jellemeztük. Az ehhez használt matematikai módszerre úgy érdemes gondolni, hogy véletlen bolyongást végzünk a definíciós gráfban. A sétát egy véletlenszerűen választott fogalomban indítjuk (hogy milyen eloszlás szerint, az, mint látni fogjuk, mindegy). A séta lépései során véletlenszerűen veszünk egyet az aktuális fogalmat definiáló fogalmak közül egyenletes eloszlással (pontosabban a multiplicitást is figyelembe véve, vagyis hogy az egyes fogalmak hányszor szerepelnek az adott definícióban). A véletlen séta határeloszlásán azt értjük, hogy hosszú idő után mekkora valószínűséggel leszünk az egyes fogalmakban (csúcsokban). Ez éppen azt fejezi ki, hogy az adott fogalom mennyire fontos az összes fogalom definiálásában, figyelembe véve azok fontosságát is.

A határeloszlás pontosan akkor egyértelmű (vagyis független a kiindulási eloszlástól), ha a gráf egyetlen erősen összefüggő komponensből áll. Láttuk, hogy ez nálunk nem áll fent. Nem erősen összefüggő irányított gráf csúcsainak az előbb leírt súlyozására szolgál a Page Rank. A Page Rank mögött szemléletesen egy olyan séta van, amely során az aktuális csúcsból 1-nél valamivel kisebb valószínűséggel továbbra is a belőle közvetlenül elérhető csúcsok egyikébe megyünk (véletlenszerűen, a multiplicitást is figyelembe véve), de kis valószínűséggel bármelyik csúcsba mehetünk. Az átmenetvalószínűségekre nézve ez azt jelenti, hogy ha az eredeti sétában P(i,j) valószínűséggel megyünk a j-edik csúcsba, feltéve, hogy aktuálisan az i-edikben vagyunk, az új sétában ugyanez az átmenetvalószínűség

$$P_d(i,j) = \frac{1-d}{n} + dP(i,j) \qquad (1)$$

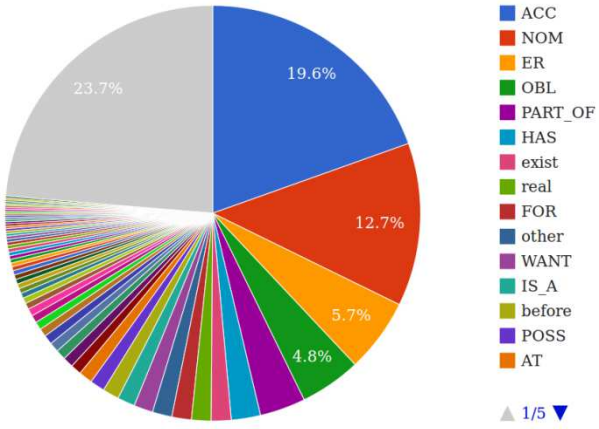
ahol d az úgynevezett csillapítási tényező (leggyakrabban d = 0.85), n pedig a csúcsok száma. d-vel 1-hez közelítve az eredeti mátrix határeloszlását közelítjük.

6 Eredmények

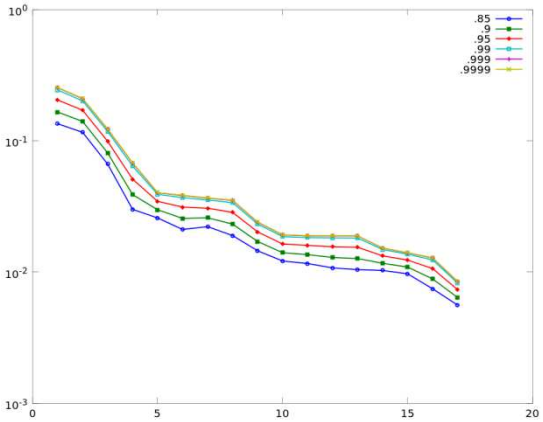
A 2. ábrán láthatjuk az eredményeket. Néhány (kevés) elem meglehetősen nagy súlyt kapott, nagyon sok pedig nagyon kicsit. A 2.b ábra a 17 legfontosabb fogalom Page Rankjét mutatja logaritmikus skálán. Itt a vízszintes tengelyen a fontosság szerinti rangsor van (d = 0.1), a függőleges, logaritmikus tengelyen pedig a megfelelő rangsorú elem Page Rank értéke. A két legfontosabb elem két mélyeset, összhangban azzal az intuitív elképzeléssel, hogy egy igei szerkezet megértésben jelentős szerepet játszik a vonzatainak megértése. Kiemelnék még két darab kétváltozós predikátumot: a jelentésreprezentációból ismerős `PART_OF`-ot és a célt (az emberhez kapcsolódó dolgok célját) kifejező `FOR`-t. Ezek az eredmények azt ígérik, hogy ahhoz, hogy helyes következtetéseket tudjon levonni egy számítógép (mesterséges intelligencia), elsősorban a rangsor elején levő elemeket kell jól kezelnie.

A 16 legfontosabb elem Page Rankje nem nagyon függ a csillapítási tényezőtől. Egyetlen kivétel a birtoklás (`HAS`), ami a leggyakoribb definiáló fogalom (a definíciók 19%-ában szerepel), ugyanakkor neki magának nincs definíciója. A `HAS` kis csillapítás (például d = 0.85) esetén nagy Page Ranket kap, erős csillapítás esetén pedig kicsit.

Az definiáló szókincs jellemzése régi probléma a lexikológiában. Ahogy Kornai–Makrai (2012)-ban is leírtuk, a 4lang egyik célja az alapszókincs meghatározása. A most közölt kutatás tehát úgy összegezhető, mint a 4lang szótár hozzájárulása az alapszókincs számszerű jellemzéséhez, vagyis ennek a régi és sok szempontból tárgyalt készletnek egyfajta formális definiálásához.



2.a ábra. A fogalmak Page Rankje



2.b ábra. A fogalmak Page Rankje logaritmikus skálán

Köszönetnyilvánítás

Szeretném megköszönni témavezetőm, Kornai András munkáját, aki ehhez a magam és a kutatói közönség számára is izgalmas feladathoz irányított, és lényeges meglátásokkal segített. Az itt közölt kutatásban – mely a korábban elkészült szótárat vette alapul – más munkatársam nem volt. Ugyancsak köszönöm Fóris Ágota, Nemeskey Dávid, Prószéky Gábor, Vámos Tibor, Váradi Tamás és az anonim bíráló hasznos megjegyzéseit. Az esetleges hiányosságokért vagy hibákért való felelősség természetesen egyedül rám hárul. A munka az OTKA Szemantikai Alapú Nyelvtechnológia (82333) pályázatának támogatásával készült.

Irodalom

Bullon , S. 2003. *Longman dictionary of contemporary English 4.* London: Longman.

Fillmore, C. 1968. The case for case. In: Bach, E., Harms, R. (szerk.) *Universals in linguistic theory.* New York: Holt and Rinehart. 1–90.

Kornai, A. 2010. The algebra of lexical semantics. In: Ebert, C. et al. (szerk.) *Proceedings of the 11th mathematics of language workshop.* Bielefeld: Springer. 174–199.

Kornai, A. 2012. Eliminating ditransitives. In: Groote, Ph. De, Nederhof M.-J.(szerk.) *Revised and selected papers from the 15th and 16th formal grammar conferences.* Bielefeld: Springer. 243–261.

Kornai, A., Makrai, M. 2012. A 4lang fogalmi szótár. In: Tanács, A., Vincze, V. (szerk.) 2012. *IX. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged: JATEPress, 62–70.