Mélyesetek a 4lang fogalmi szótárban

Makrai Márton

MTA SZTAKI Nyelvtechnológiai Kutatócsoport
e-mail: makrai@sztaki.mta.hu

Kivonat A predikátumok (főleg igék) argumentumhelyeit mélyesetekbe soroljuk, amelyek a szemantikai és szintaktikai tulajdonságok között teremtenek összeköttetést. A 4lang esetében a mélyesetek nyelvfüggetlen általánosításokat igyekszenek megragadni, melyek segítségével egy mondat jelentésreprezentációja elkészíthető.

A számítógépes szövegértés gyakran használt eszközei a mélyesetek, melyek egyszerre szemantikai és szintaktikai alapon osztályozzák a predikátumok (főleg igék és relációs főnevek, pl. *érdeke valakinek*) argumentumait, és így beazonosítható velük, hogy az egyes frázisok melyik szemantikai argumentum szerepét játsszák. A 4lang egy általános gépi szövegértés céljával készült rendszer lexikona. A cikkben a szótár rövid bemutatása és más erőforrásokkal való összehasonlítása után (1. szakasz) azt írjuk le, hogy hogyan működnek benne a mélyesetek (2. szakasz), és ezt hogyan valósítják meg (3. szakasz).

1. A 4lang fogalmi szótár

A 4langet [8]-ban mutattuk be először, majd a szavaknak egymás definiálásában való fontosságát jellemeztük a szótár segítségével [10], illetve azt vizsgáltuk, hogy a szavak jelentésének kompozícionalitása hogyan jelenik meg egy a 4langből készített vektoros nyelvmodellben (*continuous vector space model*) [9]. Mélyeseteket [14]-ben használtunk először, de még nem a most bemutatott esetkészletet. Itt a véglegesnek szánt esetrendszert mutatjuk be, melynek tesztelése még további kutatás tárgya lehet. Röviden összefoglaljuk a szótárnak azokat a tulajdonságait, amelyek a későbbiek szempontjából fontosak: többnyelvűségét és absztrakt jellegét. Magát a szótárat és néhány kapcsolódó linket a http://hlt.sztaki.hu/resources/4lang/ címen talál az olvasó.

A 4lang tételei elvileg nyelvfüggetlenek. Az eredeti változatban négy nyelven szerepelnek a tételek (angolul, magyarul, latinul és lengyelül, innen az elnevezés), ami később kiegészült negyvenre [1]. A tételek más szempontból is absztraktabbak, mint a legtöbb hagyományos vagy gép által olvasható szótár tételei, ugyanis egy absztrakt fogalmi jelentést igyekeznek megragadni, így a szótár nem törődik a szófaji különbségekkel, és egy szóalaknak csak homonímia esetén feleltet meg több tételt. Utóbbit módszertanilag ahhoz kötjük, amikor egy szóalaknak (pl. az angol *state*) egy másik nyelven két olyan szó felel meg, amelyek között nincs különösebb jelentéskapcsolat (szinkrón szinten ilyen a magyar *állam* és *állapot*).

A 4lang tehát egy többnyelvű szótár absztrakt tételekkel. [8]-ban a Webster's Third [4], [10]-ben pedig a WordNet [11] egy-egy tételén szemléltettük ezt a különbséget. Most hadd említsük még meg a BabelNetet [13] és a VerbNetet [6]. A BabelNet a 4lang-hez hasonlóan többnyelvű, de ismét csak poliszémebb felfogású, és nem csak szótár, hanem ontológia is. A VerbNetet az teszi érdekessé a jelen cikk szempontjából, hogy – bár csak az angolról szól – eseteket (tematikus szerepeket) is használ és formalizálja a jelentést. Utóbbit szemantikai predikátumokkal teszi, Moens és Steedman [12] eseményfelbontásához (*event decomposition*) hasonló módon. Tematikus szerepből 26 van, míg a 4lang-ben, mint látni fogjuk, csak nyolc, és ebből is három lokatív. A 4lang fő újdonsága itt az, hogy míg a VerbNet a tárgyas, tárgyatlan, passzív stb. használatoknál más-más vonzatkeretet (tematikusszerep-listát) ad meg, addig a 4lang szerint egy fogalomnak csak egy reprezentációja van, és az ige által kiosztott összes mélyeset abban szerepel.

2. A mélyesetek szerepe a 4langben

Amikor egy mondat jelentésreprezentációját akarjuk kiszámítani, fel kell térképezni a predikátum–argumentum viszonyokat. Elméleti nyelvészeti szempontból itt két támaszunk van: a szelekciós megszorítások, és a tág értelemben vett felszíni esetek (nyelvenként változó módon pl. a frázisok sorrendje, esetragok és/vagy adpozíciók). Kutatócsoportunk felfogásában a szelekciós megszorításoknak a *terjedő aktiváció (spreading activation)* felel meg a szótárban, a felszíni esetekkel kapcsolatos tudást pedig közvetve a mélyesetek kódolják. A mélyesetek szempontjából fontos, hogy a 4lang minden nyelvhez egy nyelvspecifikus modullal fog kapcsolódni, ami megmondja, hogy az egyes mélyesetek az adott nyelvben mely felszíni esetekkel valósulnak meg. Ebben a cikkben a mélyesetekkel foglalkozunk, ezért az aktivációterjedést csak röviden és leegyszerűsítve vázoljuk fel.

Tekintsük az úgynevezett definíciós gráfot, melynek csúcsai a szótárban szereplő fogalmak, és két ilyen akkor van összekötve, ha valamelyik szerepel a másik definíciójában (lásd bővebben a már idézett [10]-ban), pl. a 'tej' össze van kötve a 'folyadék'-kal. Ha szeretnénk megtudni, hogy egy mondatban a *tej* szó az *iszik* szó melyik argumentumát tölti be, akkor megkeressük a két fogalom között a legrövidebb utat (élsorozatot) a gráfban. Szerencsés esetben ez áthalad a *folyadék* szón, és nagyjából megfelel a *tejet iszik* kifejezés reprezentációjának.

Térjünk most rá arra, hogy a felszíni esetekből hogyan lehet kiszámítani, hogy az egyes vonzatok melyik argumentum szerepét töltik be. Egy olyan kifejezés jelentésreprezentációját, amely egy régenst a vonzataival együtt tartalmaz (pl. igei frázis), a kompozícionalitás elve szerint a következőkből kell kiszámítanunk: a régens jelentésreprezentációja, a vonzatoké, valamint az az információ, hogy mindezek együtt milyen szerkezetet alkotnak. Az utóbbiról a 4lang esetében úgy gondoskodunk, hogy a régens (tipikusan ige) jelentésreprezentációjában feltüntetjük, hogy az egyes vonzatok jelentésreprezentációjának hova kell kerülnie. Ehhez a többargumentumú predikátumok (pl. tárgyas igék) argumentumait

meg kell tudnunk különböztetni. Ezt a vonzatok mélyesetére való hivatkozással tesszük. Módszerünk mögött az a közkeletű feltételezés húzódik, hogy az argumentumok szemantikai szerepe (pl. ágens) és szintaktikai tulajdonságai (a vonzat felszíni esete, mely mondatalternációkban vesz részt a szerkezet) között (nyelveken belül) rendszeres, és több esetben különböző nyelvekben is felbukkanó megfelelések vannak, ha nem is kivétel nélküliek.

Ahogy már írtuk, a mélyesetek nálunk csak arra szolgálnak, hogy a beazonosíthassuk, hogy melyik vonzat melyik. Ennek kapcsán nem árt talán hangsúlyozni, hogy az argumentumok mélyesetekbe való besorolása elsősorban nem szemantikai osztályozás. A számítógépes szemantikában sokszor érv lehetett két mélyeset közötti különbségtétel mellett az, ha a megfelelő argumentumok jelentése között van egy rendszeres különbség. Például Talmi a *hide/mislay, pour/spill, ...* igepárok tagjai közötti szándékosságbeli különbséget annak tulajdonítja, hogy az ágens pontosan milyen esetű. Nálunk ilyen különbségek nem indokolják új mélyeset bevezetését, hiszen a jelentés teljesen le van írva a lexikai tétel definíció részében.

A szemantikai alapú osztályozáshoz képest a másik véglet az, ahol az esetek száma nem haladhatja meg a legnagyobb argumentumszámot, amivel az igék között találkozunk. Mi erre sem törekszünk, hiszen szabályszerűségeket szeretnénk kihasználni a szemantikai szerep és a szintaktikai tulajdonságok között.

3. Az egyes viszonyok

3.1. Kevéssé tartalmas elemek

Hogyan ragadja meg a `4lang` az egyszerűbb függőségeket? Egyrészt bizonyos ragoknak, pl. a többes számnak fogalmi jelentésük van abban az értelemben, hogy annak a szerkezetnek a jelentésreprezentációjában, aminek a rag része, van egy olyan eleme, amiért a rag felel. Hasonlóak a produktív képzők és az adpozíciók is. Ezeket a viszonyokat (tő–rag, tő–képző, adpozíciós tárgy–adpozíció) már csak azért is egységesen kell kezelnünk, mert a `4lang` nyelvfüggetlen kíván lenni, és ugyanazt a szemantikai viszonyt különböző nyelvek különbözően fejezik ki, pl. azt a jelentést, ami a magyarban a birtokos személyrag fejez ki, az angolban birtokos névmás. Itt a funkcióelem reprezentációjának a tartalmasabbik elem reprezentációjában való helyét mindig a `REL` (*relációs, related*) kulcsszó képviseli, amit tágabb értelemben mélyesetnek is nevezhetünk.

3.2. Igei mélyesetek

Rátérve most már az igék argumentumaira, először is tisztáznunk kell, hogy mit értünk argumentum alatt. Csak a kötelező vonzatokat, vagy a szabad bővítményeket is? A felszíni argumentumokról beszélünk, vagy a az igének egy formális szemantikai fordításban megfelelő függvény argumentumairól? Ezekre a kérdésekre első közelítésben azt válaszoljuk, hogy azokat a felszíni argumentumokat jelenítjük meg mélyesettel egy ige definíciójában, amire a jelentés leírásához szükség van. Egy másik szempont abból adódik, hogy a `4lang` absztrakt volta miatt

nem teszünk különbséget ugyanazon igealak tárgyas (esetleg még több felszíni argumentummal rendelkező) és tárgyatlan használata között. A mélyeseteket úgy kell meghatároznunk, hogy az különböző használatokban ugyanaz a szereplő ugyanazt az esetet kapja. Ebből következik, hogy ha egy igének van tárgyas használata, akkor két szereplő mélyesetét is fel kell tüntetni. Végül árnyalja a képet, hogy amikor egy ige egy másik ige speciális eseteként definiálható, és az argumentumok is öröklődnek, akkor nem szükséges kiírni az eseteket, Például a *bite* igét `cut, INSTRUMENT tooth`-ként definiáltuk ('foggal szakít'), és a *harap* örökli a *szakít* argumentumait, ezért ezeket nem tüntettük föl.

Az igei mélyesetek megválasztásánál nem feladatunk, hogy különböző igei tövek szereplői között összhangot teremtsünk. Így például nem célunk, hogy a *János elad Péternek egy könyvet* és a *Péter vesz egy könyvet Jánostól* mondatok szereplői a két mondat esetében ugyanazokat a mélyeseteket kapják.[1] Végül nem írjuk bele az igető definícióba az úgynevezett külső szerepeket (*outer role*), vagyis azokat a lehetséges bővítményeket, amikkel egy olyan konstrukció tudja felruházni az igét, ami egész igeosztályokat (pl. mozgásigék) vagy akár minden igét érint, így a következő példákban a vastagon szedett frázisoknak megfelelő (nemlétező) argumentumpozíciót: *fest egy képet* ***valakinek***, *alszik* ***egy órát***, *átröpüli* ***az Atlanti-óceánt***[2]. A kauzációt (*úszik* → *úsztat*), képzésnek tekintjük, és így a képzőt tesszük érte felelőssé (annak ellenére, hogy az angolban zéróképzéssel történik).

AGT	383
PAT	311
REL	81
POSS	52
DAT	30
TO	17
FROM	11
AT	2

1. táblázat. Az egyes mélyesetek az őket használó szavak számával

A `4lang`ben 744 igét találunk. A mélyeseteket a 1. táblázat tartalmazza azzal a számmal együtt, hogy hány szónál fordulnak elő. A leggyakoribb mélyeset az ágens. A definíciók írásakor nagyjából problémamentesen el tudjuk dönteni, hogy egy többargumentumú igének melyik argumentuma az ágens (amit az `AGT` kulcsszó jelöl a szótárban). A `4lang`-ben második leggyakoribb mélyesetet, amit mi páciensnek neveztünk (`PAT`), sokszor csak a „szemantikailag jelöletlen" mélyesetként határozzák meg, de mivel a többi viszonylag egyértelműen beazo-

[1] Előrebocsátjuk, hogy mindkét ige esetében a magyar, (vagy ami ugyanaz, az angol) alany lesz az ágens, és a tárgy a `PAT`.

[2] A külső szerepekről lásd bővebben [15] 9. fejezetét.

nosítható, ez sem okoz problémát. A szintaxisban bevett unakkuzatív hipotézis szerint egyargumentumú igék argumentuma is lehet páciens (pl. *süllyed, fürdik*). Arról, hogy az intranzitív és tranzitív igék ágensét illetve páciensét hogyan sorolják felszíni esetbe különféle nyelvek, jó összefoglalót ad Komlósy [7]. Áttekint számos ergatív (illetve aktív és alanyjelölő) nyelvet abból a szempontból, milyen esetet kap különböző egyargumentumú igék argumentuma. A 2. táblázatban mutatja be, hogy a különböző nyelvek hol húzzák meg a határt a két eset között egyfajta aktivitási skálán. Ezek az adatok azt sugallják, hogy egy nyelvfüggetlen esetrendszerben a bináris `AGT` vs `PAT` felosztásnál finomabb különbségeket kell tennünk. Kérdés, hogy ez valóban javítana-e a rendszerünkek ezeken a nyelveken való teljeseítményén. Ezt egyelőre nem tudjuk megvizsgálni, így maradunk az egyszerűbb esetkészletnél.

	tárgyjelölő	ergatív 1.	ergatív 2.	aktív	lexikalizált aktív	alanyjelölő
Péter írja a levelet.	nom	ag	ag	ag	ag	ag
Péter ír.	nom	nom	ag	ag	ag	ag
Péter sétál.	nom	nom	nom	ag	ag/nom	ag
Péter beteg.	nom	nom	nom	nom	ag/nom	ag

tárgyjelölő	angol (eng), magyar (hun)
ergatív 1.	kabardi (kbd), avar (ava), adige (ady)
ergatív 2.	agul (agx), udi (udi)
aktív	bacbi (bbl)
lexikalizált aktív	grúz (kat), dakota (dak)
alanyjelölő	megrel (xmf), maidu (nmu)

2. táblázat. Egyargumentumú igék argumentuma különféle nyelvekben [7]. A nyelvek SIL kódját is feltüntettük.

Az ágenssel és a pácienssel lényegében a generatív szemantikai hagyományt követjük. Jobban eltérünk az előzményektől a datív (`DAT`) használatával. Az elnevezést a generatív szemantika legrégebbi szóhasználatából vesszük [5,3], és datívnak magyarítjuk, a datívusz szót fenntartva az ilyen nevű felszíni esetnek. Maga Fillmore később a datívot különbontotta experiensre, tárgyra (*Object*) és célra (*Goal*). Mi a datívot alapvetően csak legalább három felszíni argumentummal rendelkező igék esetén használjuk, más esetekben csak az ezekkel való hasonlóság alapján. A háromargumentumú igék egy része jelentésére nézve a *mond* speciális esete vagy legalábbis kommunikációról szól (*bevall, enged, parancsol, kijelent, megmagyaráz, kifejez, megtilt, hálás, köszön (valamit valakinek), bemutat, mond, mutat, esküszik*), egy másik csoport pedig az *ad* speciális esete vagy változata (*kölcsönad, átenged, bérbe ad, áldoz (istennek), ajánl, tartozik, fizet, elad, adományoz, ad, segít*). Megjegyezzük, hogy a *valaminek tart* és a

valami(lyen)nek tűnik igéknek az a vonzata, ami a magyarban datívuszos, nem datívban van, hanem a mienknél finomabb felosztás szerint [2] komplementum (Comp) lenne, amit mi a `PAT`-ba sorolunk.

A `4lang`ben három lokatív eset is van, a Fillmore-i Célnak (*Goal*) illetve Forrásnak (*Source*) megfelelő `TO` illetve `FROM`, valamint a statikus `AT`. Meglehetősen messzire mentünk az absztrakcióban: ha egy vonzat sok nyelvben olyan felszíni esetet kap, ami mozgás céljának kifejezésére is használatos (a magyarban főleg határozóragok, az angolban pedig prepozíciók), akkor célnak tekintjük. Ide sorolódtak: *képes, hozzászokik, (hozzá)ad, összeadás, meghív, csatlakozik, tesz, hasonlít*. A másik két lokatív eset a forrás (*elfogad, kölcsönöz, vesz, levág, ered, eltávolít, bérel, kivon, elvesz*) és a statikus hely is (*valahol helyezkedik el, marad*).

A már említett nyelvspecifikus modulban lehetőség van megjelölni egyes igék egyes argumentumait felszíni esetekkel, amennyiben azok nem a mélyesetükből jósolható esetet kapják (*ferde eset, quirky case*). Viszont már az angol, magyar és német alapján világos, hogy maradnak igék, amelyeknél nem lehet általánosítani. Ekkor ugyanazt a `REL` kulcsszót használjuk, mint az egy felszíni argumentumú régenseknél (*prefer to something, jobban szeret valaminél*).

3.3. Relációs főnevek

Végül arról a viszonyról beszélünk, amit a *relációs főnevek* és a hozzájuk kapcsolódó szó (pl. az *érdek* szó esetén az érdekelt) között van. A jelenség, ami miatt az *érdek* főnevet relációsnak nevezem, kettős. Egyrészt az *érdek* szó előfordulásai között a birtokjelesek aránya légyegesen nagyobb, mint más főneveknél. Másrészt, és szemantikai szempontból ez az érdekes, bárhogy akarnánk is leírni az *érdek* szó jelentését, alighanem hivatkoznánk az érdekeltre. A két szó közötti grammatikai viszony a legtöbb esetben birtokos, de az esetek körülbelül egytizedében mást találunk. Az *alkalom* és a *szükség* szavak esetén az a szereplő, amit jobb híján célnak hívunk, magyarul szublatívuszba kerül (*-rA*), angolul (*occasion, need*) pedig *for* prepozíciót kap. A relációs főnevek reprezentációjában a két szó közötti grammatikai viszony szerint a `POSS` illetve a `TO` kulcsszóval jelöljük azt a helyet, ahova a kapcsolódó szó (az érdekelt illetve a cél) reprezentációja kerül. A `TO` ugyanaz az absztrakt cél, amivel már az igéknél is találkoztunk. A mélyesetek közvetítésével így a nyelvi viszony segít megtalálni a két dolog (az érdekelt és az érdek, illetve az alkalom és a cél) közötti szemantikai viszonyt. Nem foglalkozunk azokkal a relációs főnevekkel, amelyek produktívan vannak képezve valamely igéből, ugyanis ezeket a `4lang` nem különbözteti meg attól az igétől, amelyből képeztük őket.

4. Összegzés

Bemutattuk, hogyan működhetnek a mélyesetek egy olyan szabadon elérhető gépi szövegértő erőforrásban, amely a mélyeseteket közvetlenül az egyes nyelvek szavainál absztraktabb nyelvfüggetlen fogalmakhoz rendeli. A kutatás következő lépése a mélyesetek működésének tesztelése egy gépi szövegértési feladatban.

Az is későbbi kutatás témája, hogy egy tágabb szókincsbe tartozó igék szemantikai argumentumainak hogyan lehet mélyesetet tulajdonítani az őket definiáló szavaknak a szótárban levő emberi nyelvű definíciója segítségével. Végül távlati célként megemlítjük a generálást, ami úgy is felfogható, hogy adott egy r jelentésreprezentáció, és azok közül a (nem feltétlenül grammatikus) szósorozatok közül, amelyekhez a jelenlegi rendszer az r reprezentációt rendel, ki kell választani a leggrammatikusabbat, és az információs helyzetnek legjobban megfelelőt.

Köszönetnyilvánítás

Köszönöm témavezetőm, Kornai András munkáját, aki kitűzte a kutatási irányt, és fontos ötletekkel segített. Az esetkészlettel kapcsolatban fontos döntéseket a kutatócsoporton belül elsősorban Nemeskey Dávidd al hoztunk meg. Köszönöm Naszádi Katának a beszélgetéseket, amelyektől sokat tisztult a mélyesetekről való képem.

Hivatkozások

1. Judit Ács, Katalin Pajkossy, and András Kornai. Building basic vocabulary across 40 languages. In *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, pages 52–58, Sofia, Bulgaria, August 2013. ACL.
2. W. L. Chafe. *Meaning and the structure of language*. Chicago: University Press, 1970.
3. Charles Fillmore. The case for case. In E. Bach and R. Harms, editors, *Universals in Linguistic Theory*, pages 1–90. Holt and Rinehart, New York, 1968.
4. Philip Babcock Gove, editor. *Webster's Third New International Dictionary of the English Language, Unabridged*. G. & C. Merriam, 1961.
5. T. Heringer. Wertigkeiten und nullwertige verben im deutschen. *Zeitschrift für Deutsche Spache*, 23:13–34, 1967.
6. Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40, 2008.
7. Andás Komlósy. Deep structure cases reinterpreted. In Ferenc Kiefer, editor, *Hungarian General Linguistics*, pages 351–385. John Benjamins. Amsterdam–Philadelphia, 1982.
8. András Kornai and Márton Makrai. A 4lang fogalmi szótár. In Attila Tanács and Veronika Vincze, editors, *IX. Magyar Számítógépes Nyelvészeti Konferencia*, pages 62–70, 2013.
9. Márton Makrai, Dávid Márk Nemeskey, and András Kornai. Applicative structure in vector space models. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 59–63, Sofia, Bulgaria, August 2013. ACL.
10. Márton Makrai. Fogalmak fontossága a definíciós gráf vizsgálatával [importance of concepts based on the analysis of the definition graph]. In Tamás Váradi, editor, *VII. Alkalmazott Nyelvészeti Doktoranduszkonferencia*. MTA Nyelvtudományi Intézet Budapest, 2013.

11. George A. Miller. Wordnet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.
12. Marc Moens and Mark Steedman. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28, June 1988.
13. Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.
14. Dávid Nemeskey, Gábor Recski, and Attila Zséder. Miből lesz a robot MÁV-pénztáros? In *IX. Magyar Számitógépes Nyelvészeti Konferencia [Ninth Conference on Hungarian Computational Linguistics]*, 2012.
15. Harold L Somers. *Valency and case in computational linguistics*. Edinburgh University Press, 1987.