A 4lang fogalmi szótár

Kornai András és Makrai Márton

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

0. Bevezetés

A 4lang fogalmi szótár három, a számítógépes nyelvészeti és pszicholingvisztikai alkalmazások számára fontos célt szolgál. Egyrészt feladata az *alapszókincs* meghatározása, ezzel a cikk első részében foglalkozunk. Másrészt feladata a *definíciós* tevékenység támogatása, ezt a cikk második, szótárunkat a lexikográfia elméleti keretei közt elhelyező része tárgyalja. Végül célunknak tekintjük a szövegmegértési feladatok (kérdés-megválaszolás, információ-visszakeresés, gépi fordítás) támogatását is, ezzel a cikk harmadik, kitekintő része foglalkozik.

Konceptuális szótárunk lemmái többnyelvűek. Jelenleg a magyaron kívül angol, latin és lengyel, innen a 4lang elnevezés, de hosszabb távon ennek 40 nyelvre való kiterjesztését tervezzük automatikus és fél-automatikus módszerekkel. Egy tipikus lemma így néz ki:

`102 átenged V pass concedo przepuścić : LET[DAT HAS ACC]`

Mint látható, a definíció nem általában az összes *átenged* vagy *pass* írásképű szóra vonatkozik, hanem csupán a magyarban a *vki vkinek vmit* vonzatkerettel egyértelműsített fogalommal. Ha úgy tetszik, az *A vert hadsereg átengedte a várost az ellenfélnek* 'concedo' típusú mondatokat e definícióval előnyben részesítjük az *Az üveg átengedi a fényt* 'transmitto' típusú mondatokkal szemben – hogy mikor melyik jelentést választjuk, és milyen elvek alapján, arról majd a 2. részben lesz szó. Itt tárgyaljuk majd a definíciós (az írott változatban a : után eső) részt is – e bevezetéshez elég annyi, hogy ezek a definíciók egy olyan formális modell elemei, melynek megvan a saját belső szintaxisa és szemantikája.

Mitől fogalmi szótár a 4lang, mit jelent számunkra a 'fogalom'? Ideákról, az emberek fejében megjelenő konceptumokról van szó. Nyelvfilozófiai tekintetben a 4lang gyökerei Platón ideaelméletétől a skolasztikus fogalomfelfogáson át Lockeig (1689) és Fregéig (1879) vezethetők vissza. Formális modellekkel dolgozunk, de egyáltalán nem utasítjuk el azt a pszichologizmust, ami a modern generatív elméletben 'kognitív szemantika' néven ismert iskolát jellemzi (Jackendoff 1972, Lakoff 1980, Wierzbicka 1985, Talmy 1988, Langacker 1987, 1991). Éppen ellenkezőleg, a formális modelleknek valós, legalábbis felfogásunk szerint valós, tárgya van, t.i. az emberek fejében található ideák.

Közismert, hogy a nyelvtudomány egy fontos szakaszát, a késői strukturalizmustól a generativizmus fellépéséig, áthatotta egyfajta szélsőséges behaviourizmus, mely szerint az emberek fejében vagy nincsenek fogalmak, vagy ha lennének

is, ezek teljességgel megismerhetetlen és szubjektív dolgok melyeknek a tudományban nincs helye. Roy Harris (1980, 1981, 1987) több kötetet szánt annak az általa *telementáció*-nak nevezett elméletnek a bírálatára, mely szerint az emberek fejében gondolatok vannak, és a nyelv ezeket közvetíti. Úgy véljük, hogy az ilyen és hasonló (neo)behaviourista bírálatokat nem további spekulációval lehet a leghatékonyabban cáfolni, hanem olyan számítógépes modellek építésével, amelyek hagyományosan a nyelvi megértés körébe sorolt tevékenységre képesek. Az ilyen modellek ősképe Leibniz *calculus ratiocinator*a, mai megfelelői pedig az olyan kérdés-megválaszoló algoritmusok mint az IBM Watson rendszere – aki működés közben lát egy ilyet, annak semmi kétsége nem marad afelől, hogy a kérdezőtől ideák jutnak el, nyelvi úton, a befogadóhoz.

1. Az alapszókincs

Már Leibnizet is erősen foglalkoztatta, hogy a hagyományos szótárakban az egyes fogalmak definíciója gyakran körkörös: az első angol szótár, Cawdrey (1604) a *heathen*-t úgy definiálja mint 'gentile', a *gentile*-t pedig mint 'heathen'. Ezt írja (idézi Wierzbicka 1985):

> Tegyük fel, hogy kapsz tőlem egy szép summát azzal, hogy Titiustól veheted át; Titius Caiushoz küld; és Caius Maeviushoz; ha mindig így küldözgetnek az egyik embertől a másikig, soha nem kapsz kézbe semmit.

Egy lehetséges kiút egy olyan alapszókincs megadása, hogy minden más szó már ezek segítségével definiálható. Erre sok próbálkozás volt – a korai elképzelésekről remek áttekintést ad Eco (1998). A modern kísérletek közül a legfontosabb az Ogden (1930) által bevezetett Basic English, mely mindössze 850 szóból áll. Ezt használja, legalábbis ezt igyekszik használni (Yasseri et al 2012) a Wikipédia "egyszerű" kiadása (simple.wikipedia.org) is. A nyelvészetben igen jól ismert Swadesh (1950) lista (255 szó) eredetileg nem alapszókincsnek, hanem glottokronológiai vizsgálatokhoz készült, de miután ennél a feladatnál alapvető, hogy a listába vett szavak minden nyelvben előforduljanak, ez a széleskörű elterjedtség már önmagában is biztosítéka annak, hogy a Swadesh lista szavai az alapszókincsből kerüljenek ki.

Az első modern lexikográfiai elveken alapuló szótár, ami az alapszókincs elvét következetesen végigvitte, a Longman Dictionary of Contemporary English (LDOCE, ld. Boguraev and Briscoe 1989) volt, és a 4lang gerincét is az itt használt Longman Defining Vocabulary (LDV, mintegy 2000 szó és kötött morféma) adja. Ezt egészítettük ki néhány olyan klasszikus listával, mint a Diederich (1939) által összeállított 300 leggyakoribb latin szó listája, Whitney (1845) szanszkrit gyöklistája, és több magyar illetve angol gyakorisági vizsgálat leggyakoribb szavai. Természetesen nincs szó arról, hogy az LDOCE definiensei kizárólag az LDV elemeit tartalmazzák, mert a szótár alkotói igen gyakran éltek az indirekt definíciós módszerrel, pl. *Saturn: the* planet *which is 6th in order from the sun and is surrounded by large rings.* De mindaddig, amíg a kiemelt elemek már az LDV által is definiálva vannak, esetünkben *planet: a large body in space that moves*

round a star, esp. round the sun, addig a körkörösség veszélye nem áll fenn, hiszen a második definíciót az elsőbe helyettesítve azt nyerjük: *Saturn: the large body in space that moves round the sun and is the 6th such large body, and is surrounded by large rings* – ez kétségkívül körülményesebb, de ugyanazt jelenti.

Sajnos nincs szó arról, hogy az LDV már önmagában alkalmas lenne fogalmi szótárnak, hiszen ehhez garantálni kell, hogy a szavaknak ugyanabban az értelemben (például *round* nem 'kerek' hanem 'körbe') forduljanak elő a definiendum mint a definiens oldalon. Garantálni kellett azt is, hogy a szavak minden előforduló kombinációja (pl. *round + up* 'összeterelés, razzia') is definiálásra kerüljön minden olyan esetben, ha definiensben is előfordul. Ez az eset nem is annyira az igekötős igéknél (*phrasal verb*) mint az egyszavas morfológiai összetételeknél (pl. az *-er, -ist* alkotta nomen agentiseknél) fordul elő gyakran. Végső soron a teljes rendszer körmentességét csak a definíciók formális nyelvi eszközökkel való megragadása és gépi elemző építése tette lehetővé.

Külön hangsúlyozzuk, hogy a cél nem a teljes körmentesség, hanem csupán az *uroborosz tulajdonság*, tehát az, hogy definiendán kívüli elem ne legyen egyetlen definiensben sem. Az természetesen elképzelhető, hogy vannak olyan elemek, amelyeket primitíveknek kell tekintenünk (nincs hozzájuk definiens) illetve olyan párok vagy n-esek, melyek kikerülhetetlenek egymás definíciójában: a *férfi*-t nehéz a *nő*-től, a *nő*-t pedig nehéz *férfi*-től függetlenül definiálni. Lássunk néhány összetettebb példát. A *fegyenc* olyan ember, akit fegyőrök fegyházban tartanak, a *fegyőr* pedig olyan, aki fegyenceket tart fegyházban. Mi a *fegyház?* Olyan hely, ahol a fegyőrök fegyenceket tartanak. A három szó egyike helyre, a másik kettő személyre utal, de mind a három csupán ebben a konstrukcióban nyeri el az értelmét. Hasonlóképpen, mi a *tojás,* ha nem az amit a tojó tojt, és mi a tojó, ha nem az, ami tojást tojik? A konceptuális szótárnak nem feladata eldönteni, hogy melyik volt előbb.

Elvben fogalmak bármelyik L listájából kiindulhatnánk, és vizsgálhatnánk hogy ezek definíciójában mely $D(L)$ fogalmak szerepelnek. Az így nyert listát tovább vizsgálva jutunk a $D(D(L)), D(D(D(L))), \ldots$ listákhoz, és azt állítjuk, hogy a folyamat már néhány lépés után konvergál, és a (nyilvánvalóan uroborosz tulajdonságú) fixpont belül marad a 4lang keretein. Tulajdonképpen mindegy is, hogy az egy nyelv szókincsében leggyakoribb szavak jelölte fogalmakból, a legtöbb nyelvben előforduló fogalmakból, a nyelvelsajátítás során legkorábban megjelenő szavakból, a diakrón nyelvfejlődésben legkorábban megjelenő szavakból, vagy akár egy teljesnek szánt szótári listából dolgozunk: definíciós módszereink garantálják, hogy a 4lang körén kívül eső elemre soha nem lesz szükség

Az alapszókincs tehát a minimális uroborosz tulajdonságú fogalomlista, amely tartalmazhat primitíveket (ilyenek lehetnek pl. a *toj* vagy a *fegy* gyökök), de akár a *függőleges* szó is, melynek definícióját nem újabb nyelvi elemek, hanem a fejünkbe eleve beépített vesztibuláris rendszer adja meg. Hangsúlyozzuk, hogy a primitívségnek nem feltétele a monomorfémikus nyelvi alak, hiszen fogalmi szótárról van szó, a nyelvi alakok csupán adatbázis-kulcsként szolgálnak. Ezek közül is kitüntetünk egyet, a *nyomtatási nevet* (*printname*), amellyel az elemre írásban is és a szoftverben is hivatkozni lehet. Ez lehetne akár az elem sorszá-

ma is, de a programhibák javítását nagyban megkönnyíti, ha mnemonikus értéke van, ezért a kiinduló példánk nyomtatási neve nem *102* hanem `pass`. A leírásban történeti okok miatt az első helyen a magyar kulcs szerepel, de a rendszer web 2.0 alapú kiterjesztésébe és javításába igencsak nehéz lenne más anyanyelvűeket bevonnunk ha az azonosítók magyarul lennének. (A cikkben vegyesen hozunk magyar és angol példákat is.)

Már az LDV is túllépett a hagyományos szólista-felfogáson annyiban, hogy nemcsak szavakat (szabad morfémákat), hanem kötött morfémákat is tartalmazott. A 4lang is tartalmaz kötött morfémákat (mind affixumokat mind gyököket), de a fogalmi rendszer teljességéhez hozzátartoznak azok a szabályok is, melyekkel a morfológiai összetétel során kialakuló jelentést is származtatni tudjuk az összetevők jelentéséből.

2. A fogalmak és a szavak viszonya

A fogalmi szótár központi eleme a definíció. Ezt támogatja, amennyiben ez egyáltalán szükséges, a grammatikai (mind feno- mind tektogrammatikai, ld. Curry 1961) típusra vonatkozó információ. A fenogrammatikai információt a szófajban tartjuk számon, a tektogrammatikai (argumentum-struktúrára vonatkozó) információ pedig a definícióban előforduló mélyesetekből lesz kiolvasható. De a hagyományos értelmező szótáraknak van számos olyan eleme is, amivel a fogalmi szótár eleve nem foglalkozik. A legfontosabb ezek közül a fonológiai információ, amely a ritka hangfestő/hangutánzó esetektől eltekintve a fogalom megértéséhez semmivel nem visz közelebb, de ugyanez vonatkozik általában a morfológiai/morfoszintaktikai információra is. Az angol *go* és *walk* szavak által jelölt, egyébként igen hasonlatos, fogalmak megértéséhez nem visz közelebb az az ismeret, hogy az előbbinek rendhagyó a múlt ideje de az utóbbinak nem. Nem foglalkozunk a szavak etimológiájával sem, bár ez gyakran segíti a megértést, de úgy véljük, hogy a nyelvelsajátítónak etimológiai információ tipikusan nem áll rendelkezésére, tehát egy ilyeneken alapuló rendszertől semmiféle kognitív realitást nem várhatunk. Nem foglalkozunk a szavak stiláris értékével sem, mert a fogalmi szótár számára elégséges ha a *kutya* definiálásra kerül, az *eb* már használhatja ugyanezt a definíciót. Megjegyezzük, hogy az általunk figyelmen kívül hagyott szótári információk az átlagos szócikk kevesebb, mint 10%-át teszik ki, akár bitben, akár nyomtatott karakterben számolva.

A definíció célja a fogalmak közti belső kapcsolatok rögzítése. Amikor a kapcsolat csupán történeti, mint pl. *bishop* 'püspök' illetve *bishop* 'futó (sakkfigura)' akkor az értelmező szótárakban szokásos módon alsó indexekkel különböztetjük meg a lemmákat. Az ilyesfajta tiszta homonímiák elkülönítése a poliszémiától már a hagyományos lexikográfiában is sok fejtörést okoz, és a nehézségek alól természetesen a fogalmi szótár sem tudja teljesen kivonni magát. Szerencsére a fogalmi szótárnál érvényesíteni tudunk több olyan tényezőt, ami a feladatot lényegesen megkönnyíti. Az első a rugalmasabb szófajkezelés: ezt lentebb a *fagy* ige ill. *fagy* főnév példáján fogjuk illusztrálni.

A második tényező módszertani. Kirsner (1993) két élesen szemben álló megközelítésről ír: a *poliszemikus* felfogás igyekszik a szavak jelentéseit maximálisan elkülöníteni, pl. $bachelor_1$ 'nőtlen felnőtt férfi', $bachelor_2$ 'pár nélküli fóka', $bachelor_3$ 'más lovag zászlaja alatt szolgáló lovag', és $bachelor_4$ 'BA vagy BSc fokozattal rendelkező személy'. A *monoszemikus* megközelítés (melyet Kirsner *Saussure*-i és *Columbia School* megközelítésnek is nevez) egy általános, absztrakt jelentést tételez, mely alá esetünkben legalább az első három aljelentés besorolható: 'tipikus férfiszerepben kielégítetlen'. A 4lang a monoszemikus megközelítést követi, ennek filozófiai alapjairól ld. Ruhl (1989).

A poliszémia minimalizálásának van még egy igen fontos módszertani indoka, amely véleményünk szerint nemcsak a 4lang, hanem minden fogalmi szótár lényegéből ered. Egy ilyen szótár célja kifejezetten a szavak egymás közti viszonyainak, nem pedig a szavak és a világ dolgai közti viszonyok feltárása. Élesen el kell különíteni a *lexikai* és az *enciklopédikus* információt, alapjában ugyanazon kritériumok mentén amelyekkel a filozófia az analitikus és a szintetikus kijelentéseket választja szét. Az értelmező szótári gyakorlatban az enciklopédikus információ gyakran keveredik a lexikaival: példaképp álljon it a *potash* 'hamuzsír, kálisó' definíciója a Webster's Third-böl:

> 1a: potassium carbonate, esp. that obtained in colored impure form by leaching wood ashes, evaporating the lye usu. in an iron pot, and calcinating the residue – compare pearl ash. b: potassium hydroxide. 2a : potassium oxide K_2O in combined form as determined by analysis (as of fertilizers) ⟨ soluble ∼ ⟩ b: potassium – not used systematically ⟨ ∼ salts ⟩ ⟨ sulfate of ∼ ⟩ 3: any of several potassium salts (as potassium chloride or potassium sulfate) often occurring naturally and used esp. in agriculture and industry ⟨ ∼ deposits ⟩ ⟨ ∼ fertilizers ⟩

Jól látható, ahogy az enciklopédikus tudás rögzítésével a szótáríró magának csinálja a poliszémiát. Az LDOCE felfogása szerint itt egyáltalán nincs szó többértelműségről:

> any of various salts of potassium, used esp. in farming to feed the soil, and in making soap, strong glass, and various chemical compounds

és a 4lang, ha a hamuzsírt az alapszókincs részének tekintené, akkor még tovább menne az absztrakcióban:

```
salt, HAS potassium
```

A modern tudásreprezentációs sémák erősen hajlanak a tudományközpontúság felé, pl. Kripke (1972) a vizet mint H_2O-t definiálja. Történetileg azonban a nyelvi tények megelőzik a tudományos ismereteket, előbb volt a hamuzsír mint a kálium. Nincs semmi okunk azt feltételezni, hogy egy minden tekintetben kielégítő biológia definíció hiányában az emberek nem tudják mondjuk a *kutya* szót használni, vagy hogy Berzelius előtt a *víz* mást jelentett mint ma. Mivel a fogalmi szótár célja nem az egyes fogalmak meghatározása, hanem ezek rendszerének feltárása, a *víz* definíciója nem H_2O, hanem

```
2622 víz N water aqua woda: liquid, NOTHAS colour, NOTHAS
taste, NOTHAS smell, life NEED
```

tehát csupa olyasmi, amit az emberek évezredek óta tudnak (és amiknek a modern tudomány akár ellent is mondhat). Ahol mégis a tudományok által definiált dolgokról van szó (ilyen pl. a *kálium* aminek egyszerűen nincs hétköznapi definíciója) ott egy külső enciklopédiára, konkrétan a Wikipédiára mutató kereszthivatkozásokat használunk:

```
potassium : element, @http://en.wikipedia.org/wiki/Potassium
```

Ugyanilyen kereszthivatkozásokat használunk ott is, ahol a lexikográfiai gyakorlat illusztrációkkal dolgozik – ez is meglehetősen ritka eset, pl. az angolszász lexikográfia alapműve, az Oxford English Dictionary egyáltalán nem használ illusztrációkat, a Websters Third pedig a szócikkek kevesebb mint fél százalékánál.

A definíciók szintaxisa arra a feltevésre épít (Kornai 2011), hogy a primitívek listája egyáltalán nem kell, hogy kettőnél több argumentumú (ditranzitív vagy magasabb aritású) elemeket tartalmazzon, mert ezeket mindig lehet egyszerűbb aritásúakkal definiálni. Jó példa a `give`, aminek a definíciója 'cause to have', egész pontosan `CAUSE[DAT HAS ACC]` – a rendszerbe beépített redundanciaszabály szerint a `CAUSE` alanya, mint minden tranzitív predikátumé, nominatívuszi. Az alapfogalmak túlnyomó része intranzitív, ilyenek a köz- és tulajdonnevek (kivéve természetesen a relációs főneveket), a melléknevek, és az intranzitív igék is. A tranzitív elemeket az írott változatban csupa nagybetűvel jelöljük. Az implementáció alapját adó gépek (machine, definícióját ld. Eilenberg 1974) kétféle változatát használjuk: egy- illetve kétpartíciósat (erről bővebben ld. Kornai 2010), attól függően, hogy az elemet intranzitívnak vagy tranzitívnak tekintjük.

Az intranzitív elemeket mint a rájuk analitikusan jellemző predikátumok konjunkcióját definiáljuk, pl. `488 düh N anger furor gniew: feeling, bad, strong, aggressive`. Annyiban Arisztotelészt és a skolasztikus hagyományt követjük, hogy a definíciónak az esszenciát kell megragadnia, de abban eltérünk a hagyománytól, hogy mi a düh *szó*, a düh *fogalom* jelentését, nem pedig a világban található reális düh lényegét próbáljuk megragadni. Ez utóbbi nyilván valami hormonszint-változással függ össze, de ezt mi enciklopédikus ténynek tekintjük, és mint ilyet figyelmen kívül is hagyjuk. Ebből adódik a 4lang egy fontos tulajdonsága: számunkra a *dobermann* és a *pincsi* definíciója egyaránt `dog`.

Kevesebb, mint harminc primitív tranzitív elemünk van, ezek között a legfontosabbak grammatikai jellegűek. A legnagyobb csoport a mélyesetek `NOM, ACC, DAT, ...`, de a melléknevek fokozásánál elkerülhetetlen az `ER`, és a főnevek birtoklásánál kikerülhetetlen egy `HAS` alak. A tisztán konceptuális binárisok közt a leggyakrabban az `AT` szerepel definiensben, ebben a monoszemikus felfogásnak megfelelően együtt szerepel az időbeli és a térbeli összekapcsolódás. A tranzitívnál bonyolultabb argumentumstruktúra megragadásának eszköze a tranzitív relációk egymásba ágyazása, pl. `1846 öl V kill interficio zabijać: CAUSE[ACC[die]]`. Ebben a tekintetben a 4lang a generatív szemantika definíciós módszereit követi, a különbség elsősorban a változók és a változókötés mechanizmusának sajátos, gépeken alapuló megvalósításában áll.

Térjünk most vissza az olyan többszófajú elemek problémájára mint az angol *divorce* vagy a magyar *fagy*. Felfogásunk szerint ilyenkor az igénél és a főnévnél ugyanarról a fogalomról van szó, t.i. arról a folyamatról, amiben a víz szilárd lesz, vagy ennek okáról: definíciós nyelvünkön `cold CAUSE, before[liquid], after[solid,<ice>]`. A természetes nyelv egy sajátos jellemzője, hogy az okot és az okozatot ilyenkor nem szemantikai, hanem fenogrammatikai eszközökkel különíti el. A tökéletes filozófiai nyelv kialakítására törekvő filozófusokat, pl. Francis Bacont, ez és a többi *idola fori* nagyon zavarta, de véleményünk szerint a szemantika a nyelvtudomány része, és mint ilyen deskriptív, nem pedig normatív módszertannal dolgozik.

A 4lang meghoz számos olyan technológiai döntést, amelyeket minden fogalmi szótárnak meg kell hoznia, de nem feltétlenül úgy, ahogy ezt mi tesszük. Ilyen az alapértelmezett (*default*) értékek konzekvens használata: az előző példánál maradva a *fagy* eredménye alapértelmezésben a jég, bár természetesen nagy hidegben az alkohol, a paraffin, de még a hőmérő higanya is megfagy. A szótárban a default értékeket `< >` jelöli. Egyedi döntés az is, hogy a *before* és *after* elemek egyváltozósak, hiszen másik változójukat úgyis a cselekvés idejéhez kellene kötnünk. Végül ugyanilyen döntés az is, hogy kikerültük az uniform Boole-jellegű negációt, helyette külön primitívnek véve a `NOTHAS, NOTAT` és hasonló negatív relációkat: pl. a *kígyó* definiáló tulajdonsága a `NOTHAS leg`, a *lélek*-nek a `NOTHAS material`, a *lop*-nak pedig a `NOTHAS right`. Van természetesen negációs primitív (intranzitív) elem, sőt többféle is van, ezek közül legfontosabb a `lack` amely normálisan (alapértelmezésben) meglévő elem hiányát jelzi: például a *beteg* `lack(health)`, ami több mint a `NOTHAS health` hiszen nem csak arról van szó, hogy nincs neki, hanem egyben arról is, hogy kellene lennie, míg ez utóbbi következtetést pl. a kígyó lábáról nem kívánjuk levonni.

3. Alkalmazások

A 4lang adja az alapját több olyan rendszernek is, melyeket munkacsoportunk már a gyakoratban is bemutatott: ilyen a SHRDLU 2.0 (Kutatók Éjszakája 2011), az Elvira-asszisztens (Edinburgh 2012), és a robotpénztáros (Kutatók Éjszakája 2012). Mint minden gyakorlatban működő rendszernél, itt is szükség van interfészekre, amik a rendszeren kívüli komponensek (a kockarakosgató robot, a www.elvira.hu weblap, illetve a pénztári adatbázis) meghajtására alkalmasak. A rendszer egésze tehát képes az ilyen és hasonló külső komponensekkel kapcsolatos enciklopédikus tudás megragadására is, de ezt formailag is eltérő, nem gépeken, hanem attribútum-érték mátrixokon (AVM) alapuló mechanizmussal teszi.

Tekintsük például az Elvira-asszisztenst, amely a www.elvira.hu-ról azt tudja, hogy ha három attribútum (dátum, kiindulás, cél) közül legalább a kiindulás és a cél már ki vannak töltve, akkor az ezekből kialakított queryt `http put` segítségével elküldi a www.elvira.hu-ra. A természetes nyelvi rendszer feladata kettős: egy magyarul megfogalmazott kérdésből, pl. *Mikor megy holnapután vonat Szegedre a Nyugatiból?* észre kell vennie, hogy ez egy olyan kérdés, amit az

Elvira meg tud válaszolni, másrészt hogy ki tudja választani az egyes attribútumokra vonatkozó értékeket: dátum: holnapután, kiindulás: Budapest Nyugati, cél: Szeged.

Ehhez a 4lang egy olyan változatára van szükség, ami a *holnap* mellett (ez benne van az alapszókincsben) tartalmazza a *holnapután* szót is, és persze a *vonat* szót is. Felhasználásra kerül a szótárnak néhány olyan eleme is, ami a példamondatban ugyan nem szerepel. de kikerülhetetlen közbenső kapocs az Elvirához: ilyen elsősorban az *Elvira* szó, ami definíciója szerint `vonat, menetrend` és enciklopédikus részében tartalmazza a fentebb leírt háromelemű AVM-et. Nyilvánvaló, hogy a rendszer csak akkor tudja hívni az Elvirát, ha tudja hogy van ilyen. A felhasználónak viszont nem kell ezt tudnia, kiinduló mondatunk nem az, hogy *Kérdezd meg az Elvirát...*

Rendszerünk logikájából adódóan szükség van még a *vonat* és a *menetrend* szavak definíciójára is, de ezekben már semmi Elvira-specifikus nincsen: a vonat számunkra `mass_transit, rail, ...` a menetrend pedig egyszerűen `mass_transit, when`. Az Elvira AVM-hez egy teljesen általános mechanizmussal, a terjedő aktivációval (*spreading activation*, ld. Quillian 1967) jutunk el az eredeti inputban szereplő *mikor* (when) illetve *vonat* (train) szavakon, illetve az inputban már nem szereplő, de ezek által aktivált *menetrend* (schedule) szón keresztül.

Köszönetnyilvánítás

A 4lang-ot használó rendszerek kialakításán legtöbbet Nemeskey Dávid, Recski Gábor, és Zséder Attila (SZTAKI) dolgoztak. A 4lang alapjait illetve az egyes definíciókat illetően számos hasznos tanácsot kaptunk még az alábbiaktól: Kálmán László (NYTI), Muntág Márton (ELTE), Rebrus Péter (NYTI), Rung András (KREA), Szakadát István (BME MOKK), Szőts Miklós (ALL), Varasdi Károly (PPKE), Vásárhelyi Dániel (ELTE). A munka az OTKA Szemantikai Alapú Nyelvtechnológia (82333) pályázatának támogatásával készült.

Hivatkozások

1. Branimir K. Boguraev and Edward J. Briscoe. *Computational Lexicography for Natural Language Processing*. Longman, 1989.
2. R. Cawdrey. *A table alphabetical of hard usual English words*. 1604.
3. Haskell B. Curry. Some logical aspects of grammatical structure. In R. Jakobson, editor, *Structure of Language and its Mathematical Aspects*, pages 56–68. American Mathematical Society, Providence, RI, 1961.
4. Paul Bernard Diederich. *The Frequency of Latin Words and Their Endings*. Illions, The University of Chicago Press, 1939.
5. Umberto Eco. *A tökéletes nyelv keresése*. Atlantisz, 1998.
6. Samuel Eilenberg. *Automata, Languages, and Machines*, volume A. Academic Press, 1974.
7. Gottlob Frege. *Begriffsschrift: eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*. L. Nebert, Halle, 1879.

8. Roy Harris. *The language-makers*. Duckworth, 1980.
9. Roy Harris. *The language myth*. Duckworth, 1981.
10. Roy Harris. *The language machine*. Duckworth, 1987.
11. Ray S. Jackendoff. *Semantic Interpretation in Generative Grammar*. MIT Press, 1972.
12. R.S. Kirsner. From meaning to message in two theories: Cognitive and saussurean views of the modern dutch demonstratives. *Conceptualizations and mental processing in language*, pages 80–114, 1993.
13. András Kornai. The algebra of lexical semantics. In Christian Ebert, Gerhard Jäger, and Jens Michaelis, editors, *Proceedings of the 11th Mathematics of Language Workshop*, LNCS 6149, pages 174–199. Springer, 2010.
14. András Kornai. Eliminating ditransitives. In *Formal Grammar*, pages 243–261, 2011.
15. Saul A. Kripke. Naming and necessity. In D. Davidson, editor, *Semantics of Natural Language*, pages 253–355. D. Reidel, Dordrecht, 1972.
16. George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago Press, 1980.
17. Ronald Langacker. *Foundations of Cognitive Grammar*, volume 1. Stanford University Press, 1987.
18. Ronald Langacker. *Foundations of Cognitive Grammar*, volume 2. Stanford University Press, 1991.
19. J. Locke. *An Essay Concerning Human Understanding*. Ward, Locke and Bowden, 1689.
20. C.K. Ogden. *Basic English: a general introduction with rules and grammar*. K. Paul, Trench, Trubner, 1944.
21. M. Ross Quillian. Semantic memory. In Minsky, editor, *Semantic information processing*, pages 227–270. MIT Press, Cambridge, 1967.
22. C. Ruhl. *On monosemy: a study in lingusitic semantics*. State University of New York Press, 1989.
23. Morris Swadesh. Salish internal relationships. *International Journal of American Linguistics*, 16:157–161, 1950.
24. L. Talmy. Force dynamics in language and cognition. *Cognitive science*, 12(1):49–100, 1988.
25. W.D. Whitney. *The Roots, Verb-forms, and Primary Derivatives of the Sanskrit Language*. Motilal Banarsidass, 1845.
26. Anna Wierzbicka. *Lexicography and conceptual analysis*. Karoma, Ann Arbor, 1985.
27. Taha Yasseri, András Kornai, and János Kertész. A practical approach to language complexity: a wikipedia case study. *PLoS ONE*, 2012.