Szakmai beszámoló az

„Analogikus általánosítási folyamatok a gyereknyelvben”

c. kutatási projekthez

Kutatásvezető: Babarczy Anna,
Budapesti Műszaki és Gazdaságtudományi Egyetem, Kognitív Tudományi Tanszék

Futamidő: 3 év

Összegzés

Elméleti összefoglaló

A lexikai tudás, vagyis a felnőtt nyelvtan által megengedett predikátum-argumentum struktúrák elsajátítását vizsgáltuk. A kutatás módszere a gyereknyelvi adatok elemzéséből nyert statisztikák összevetése különböző számítógépes tanulási mechanizmusok eredményeivel.

A CHILDES adatbázisból elérhető és a projekt keretében készített magyar gyereknyelvi korpuszokat a kutatás céljaira kialakított annotációs rendszerben elemeztük az előforduló predikátum-argumentum szerkezetek helyessége szerint. Az elemzés eredményeként sekély U-görbét kaptunk, ami arra utal, hogy a kezdeti konzervatív tanulási mechanizmust felváltja egy analogikus általánosító mechanizmus, amely átmenetileg hibákhoz vezet.

A gyerek nyelvelsajátítási mechanizmusainak szimulálására automatikus vonzatkeret-kinyerő alkalmazást hoztunk létre. Elsőként Brent által kidolgozott statisztikai gépi tanulási módszert adaptáltuk a magyar nyelvre. A tanulás a vonzatok morfológiai jegyei alapján történik annotált korpuszból. Brent módszere szigorú konzervatív tanulási algoritmus, ahol a vonzatkeretek elsajátítása kizárólag megfelelő pozitív input alapján történik, így nem kaptunk a gyereknyelvi adatokhoz hasonlítható U-görbét. Második lépésben a tanulási algoritmust úgy módosítottuk, hogy ne zárjuk ki az általánosítás illetve túláltalánosítás lehetőségét. Ez a modell közelebb áll a gyereknyelvben megfigyelt mintákhoz, de lényegesen több inputra van szükség. A cél-nyelvtan leszűkítésével eredményjavulást értünk el.

Eszközök, fejlesztések

- Gépi tanulás: saját fejlesztésű modellek.
  A modellek tanításához a Szeged Treebank (Csirik, Gyimóthy, Kis, Prószéky, Várady) és a Magyar Webcorpus-t (Halácsy et al. 2004) használtuk. Ezek morfológiai annotációját és egyértelműsítését a HunMorph gépi elemzőcsaládd al (pl. Halácsy et al. 2007) végeztük. Az morfológiai elemzés a „KR” annotációs nyelvtant használja (ennek részletes leírását ld. http://ftp.mokk.bme.hu/Language/Hungarian/Freq/Web2.2/kr_for_ldc.pdf) (Trón et al, 2006).

  A modellek értékeléséhez „gold standard” készült a felnőtt nyelvtan által

megengedett ige és vonzatkeret párokról. A Szeged Treebank és a Webcorpus korpuszokban előforduló leggyakoribb 1000 igére vonatkozó vonzatkeret nyelvtant saját fejlesztésű software segítségével készítettük. Eredményeink publikálását követően a nyelvtant és a szerkesztő programot nyilvánosságra hozzuk.

- Gyereknyelv: új korpusz.
  A gyereknyelvi elemzésekhez a nemzetközi Child Language Data Exchange System (CHILDES) adatbázisának magyar nyelvű longitudinális spontán gyereknyelvi korpuszait (MacWhinney és Réger gyűjtései), és a projekt keretében hangfelvételről átírt magyar gyereknyelvi korpuszt (Réger gyűjtése) használtuk. Az utóbbi anyagot a CHILDES formátumának megfelelően digitalizáltuk és átírtuk. Kutatási eredményeink publikációit követően a korpuszt a CHILDES adatbázison keresztül további kutatási célokra elérhetővé tesszük.

Tervezett publikációk

- A gépi modellekre fektetve a hangsúlyt, kézirat készül a 2009. decemberében tartandó VI. Magyar Számítógépes Nyelvészeti Konferencia kiadványába.
- A gyereknyelvi eredményeket a Journal of Child Language nyelvelsajátítási modellekkel foglalkozó különkiadásában tervezzük publikálni angol nyelven.

A kutatási tervtől való eltérések

- A számítógépes modell első változatának fejlesztését az eredeti munkatervvel ellentétben a gyereknyelvi korpuszok statisztikai elemzése előtt végeztük el. A témába vágó konferencia kiírások fókusza miatt nagyobb hangsúlyt fektettünk a számítógépes modellre, mint eredetileg terveztük.
- 2008. évben a vártnál magasabbak voltak a projekt személyi költségei. A kutatóasszisztens hallgatói jogviszonya megszűnt, és ezzel megnövekedtek a járulék terhek. A hiányzó összeget nagyrészt a konferenciákra szánt összegből pótoltuk.
- A kutatás lezárásának dátumát az OTKA előzetes engedélyével 2009.01.31-ről 2009.05.31-re módosítottuk. A módosítás oka az volt, hogy 2009. május 22-én a *Dubrovnik Conference on Cognitive Science: Language and the Brain* c. nemzetközi konferencián mutattuk be eredményeinket, és ennek a pályázatból való finanszírozására csak a záró dátum kitolása esetén volt lehetőség.

Az eredmények részletes ismertetése

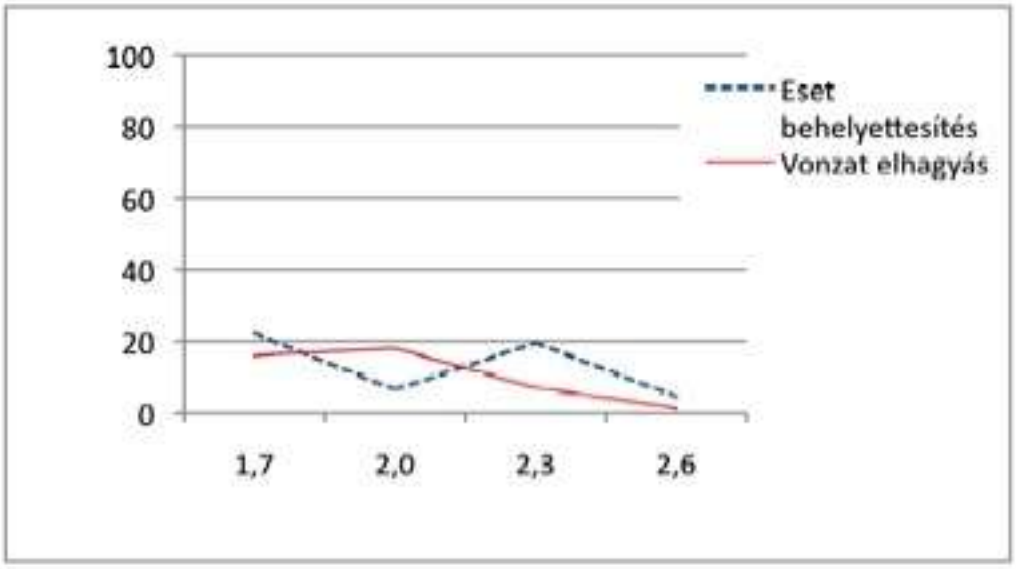
1. A lexikális tudás kérdése

Lexikális tudás elsajátítása alatt a szavak és ezek idioszinkratikus (nem általános elvekből következő) tulajdonságainak elsajátítását értjük, beleértve szemantikai és szintaktikai tulajdonságokat. A predikativ nyelvi elemek – köztük az igék – lexikális tulajdonságai közé tartozik a vonzatszerkezetük, azaz hogy milyen kategóriájú illetve morfo-szintaktikai szerkezetű bővítményekkel jelenhetnek meg a mondatban. Ez a tudás nem csak a mondatalkotás, hanem a mondatfeldolgozás szempontjából is elengedhetetlen. Például az *elad* és a *megsimogat* igék vonzatkeretének ismeretében tudjuk azt, hogy míg az alábbi (1) mondat kétértelmű (Lili szomszédja lehet a cselekvés célpont argumentuma, vagy a kutya eredeti gazdája), a (2) alatt szereplő mondat nem az (Lili szomszédja itt nem lehet argumentum).

(1) Marci eladta Lili szomszédjának a kutyáját.

(2) Marci megsimogatta Lili szomszédjának a kutyáját.

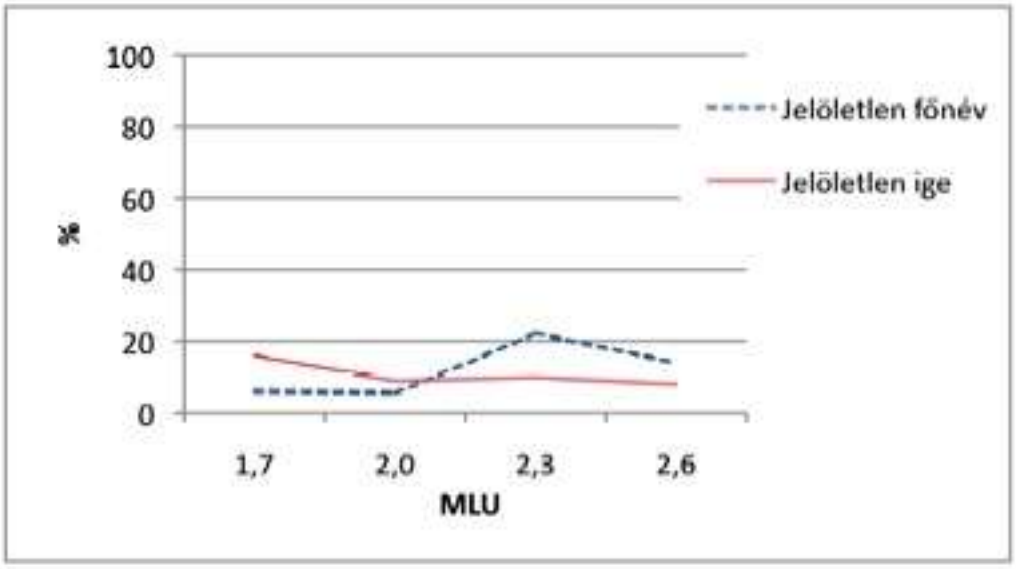
A lexikális tudás elsajátításának mechanizmusai két szempontból is érdekes kutatási téma. Egyrészt a pszicholingvisztikában fontos kérdés a nyelvi tudás ezen alapelemének fejlődése, másrészt a számítógépes nyelvfeldolgozás területén a gépi parsing rendszerek egyik fő problémája. Kutatási projektünk a gyereknyelv empirikus tapasztalataiból kiindulva próbálja a gépi nyelvfeldolgozás módszereit fejleszteni, míg a másik irányban a számítógépes modellek működésén keresztül igyekszünk fényt deríteni az empirikus tapasztalatok mögött rejlő emberi tanulási mechanizmusokra. A korai automatikus lexikonépítési kísérletekben nem számítógépes célokra készült szótárak elektronikus változatát használták nyersanyagként. Az automatikus módszerek közül ez a megközelítés áll legközelebb a kézi előállításhoz, éppen emiatt rendelkezik a nem automatikus módszer fő hátrányaival: nem elég rugalmas, és nem teszi lehetővé az automatikus bővítést, ezáltal nem vihető át más területre. A szótár használatánál robusztusabb megközelítést jelent az igei vonzatkeret-információ automatikus kinyerése nagyméretű korpuszokból. A gyereknyelvi adatok is arra utalnak, hogy az anyanyelv elsajátításakor nem az egyes igék vonzatszerkezetének egyenkénti memorizálásával épül a mentális lexikon, hanem az input statisztikai tulajdonságait felhasználva mintákat vonnak ki a gyerekek a nyelvi inputból, ami a tanulás egyes szakaszaiban hibákhoz vezethet. Amint az 1. ábrából kiderül, a gyereknyelvben előforduló vonzatkeretek nem mindig felelnek meg a célnyelvtan által elfogadott vonzatkereteknek.



1. ábra: Helytelen nem alanyi esetragok és elhagyott kötelező vonzatok aránya a korai magyar gyereknyelvben. Három gyerek spontán nyelvi produkciójának súlyozatlan átlaga. Korpusz méret: 18 644 szó.

A feladatot úgy fogalmazhatjuk meg, hogy ha adott egy *F* vonzatkeret készlet és egy *V* igehalmaz, az inputban megjelenő mondatok alapján döntsük el minden $(f, v) \in F \times V$ párról, hogy a nyelvtan szerint *f* lehet-e *v* vonzatkerete. A tanulás eredményeként megengedett ige-vonzatkeret párok alkotják a tanuló lexikonját. A gyereknyelv esetében az input a a gyerek nyelvi környezetét jelenti, a számítógépes modell pedig digitális korpuszokból tanul. A továbbiakban igei vonzatkeret alatt egyszerűen azt az információt értjük, hogy az ige bővítményei a mondatban milyen (felszíni) esetben vannak, mivel a magyar nyelvben a vonzatok szintaktikai illetve tematikai szerepét elsősorban az esetrag jelöli.

A fenti leírás feltételezi, hogy a gyerek számára is adott egy vonzatkeret készlet és egy igekészlet, és a feladata hasonlóképpen az, hogy az igékhez a megfelelő vonzatkereteket rendelje. Ezt a feltételezést az a megfigyelés támasztja alá, hogy a korai gyereknyelvet egyszavas mondatok jellemzik, igék és főnevek egyaránt, melyeket tekinthetünk predikátumok és argumentumok egyszerű megjelenítésének. A magyar (és más gazdag morfológiájú) nyelvekben a korai gyereknyelv mondatai jellemzően ragozott szavakból állnak: az igék inflexiókkal, a főnevek pedig esetragokkal jelennek meg. Természetes gyereknyelvi korpuszelemzéseink megerősítették ezt a megfigyelést: a 2. ábrán látható, hogy viszonylag kevés inflexiói-elhagyási hiba fordul elő a magyar gyereknyelvben azelőtt is, hogy az átlagos mondathossz elérné a két szót (a jóval gyakoribb morfofonológiai hibákat és rag-behelyettesítéseket itt figyelmen kívül hagyjuk).



2. ábra: A jelöletlen (esetraggal nem ellátott, nem alanyi szerepű) főnevek és a jelöletlen (személyraggal nem ellátott, nem egyesszám harmadik személyű alanyú) igék aránya a korai magyar gyereknyelvben. Három gyerek spontán nyelvi produkciójának súlyozatlan átlaga. Korpusz méret: 18 644 szó.

Feltesszük tehát, hogy a gyerek számára adott a világ eseményeinek és az azt leíró nyelvnek predikátumokba és a hozzájuk tartozó argumentumokba való szerveződése. A fenti adatokra támaszkodva feltesszük továbbá, hogy a gyerek számára ismert az esetragozás mechanizmusa. Ezek a nyelv általános törvényszerűségeiből következő tudások, melyek eredetével kutatásunk nem foglalkozott.

2. A gépi modellek

Alapelvek

Kutatásunk fő irányvonala az argumentumstruktúrák elsajátításának számítógépes modellezése volt. A vonzatkeretek gépi tanulására első megközelítésként Brent (1993) statisztikai módszerének gazdag morfológiájú nyelvekre adaptált változatát alkalmaztuk, Bár Brent módszere – a számítógépes nyelvészet fejlődési ütemét tekintve – elég réginek nevezhető, magyar vonzatkeretek azonosítására (tudomásunk szerint) ez az első alkalmazása. A magyar nyelvvel foglalkozó munkák közül a miénkhez hasonló tárgyú Sass (2006) munkája, de ez az idiomatikus, nem kompozicionális, rögzített lemmával előforduló igei szerkezetek kigyűjtését tűzi ki célul.

Röviden, Brent eljárásának az a feltételezés az alapja, hogy minden vonzatkerethez tartoznak ú.n. jegyek. Egy jegy olyan mintázat vagy formai sajátosság, aminek megjelenése egy mondatban valószínűsíti, hogy a mondatban előfordul a jegyhez tartozó igei vonzatkeret. Például, a „tárgyas ige” vonzatkerethez tartozhat a következő jegy: a mondatban pontosan egy ige van, és van benne tárgyesetű névszó. Az általunk használt jegyrendszer egyszerű reguláris kifejezésekből áll, melyek a KR morfológiai annotációs kód (Trón et al 2006) elemeire illeszkednek: egy jegy illeszkedik egy mondatra, ha a megfelelő reguláris kifejezés illeszkedik a

mondathoz tartozó morfológiai annotáció sztringre. Az 1. táblázat megjelenít egy példát.

mondat	KR annotáció
Én	NOUN<PERS<1>>
ma	ADV
már	ADV
nyertem	VERB<PAST><PERS<1>>
egy	NUM
telefont.	NOUN<CAS<ACC>>

1. táblázat: Mondat morfológiai annotációja a KR-kód felhasználásával.

A magyar ditranzitív vonzatkeret például a következő kódnak felel meg:

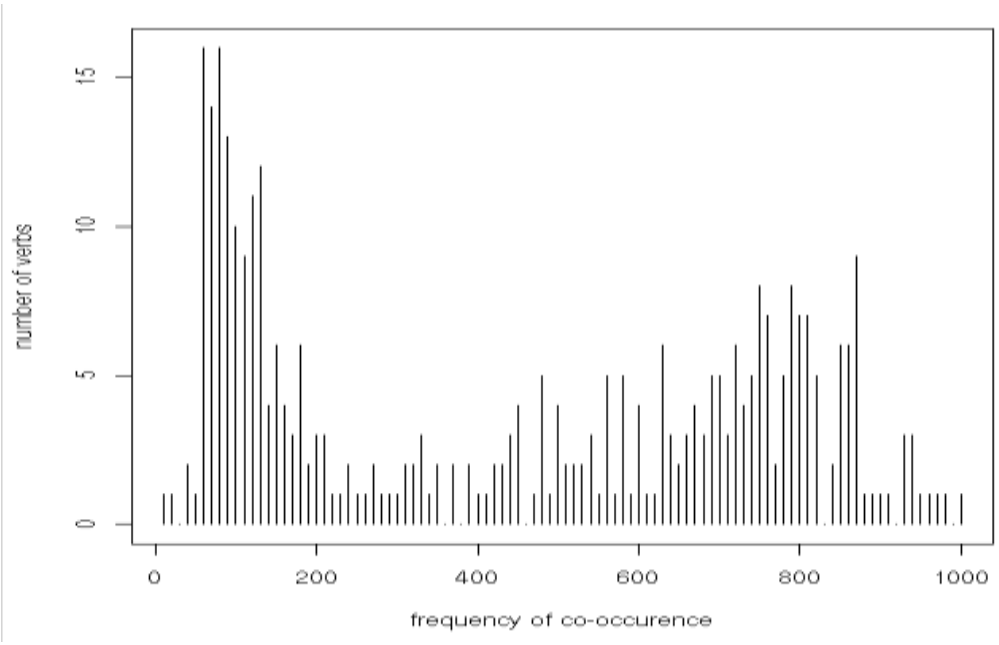
(3) (CAS<ACC>.*CAS<DAT>) | (CAS<DAT>.*CAS<ACC>).

A számítógépes modellben felhasznált jegyeket a gyereknyelvi korpuszban konzisztensen előforduló, a felnőtt nyelvtan szabályainak megfelelő argumentumszerkezetek részletei adják.

Minden jegyhez tartozik egy hibavalószínűség, ez annak a valószínűsége, hogy a jegy ugyan megjelenik egy mondatban, de a jegyhez tartozó vonzatkeret mégsem tartozik az adott predikátum megengedett vonzatkeretei közé.

Hibavalószínűségek

A hibavalószínűségek (ε) meghatározása különböző módszerekkel történhet. Elméleti szempontból az a módszer tűnt az emberi nyelvelsajátítás legjobb megközelítésének, amely a vonzatkeretek disztribúciójára épül. (Amint a 3. fejezetben látni fogjuk, végül nem ez a módszer bizonyult a legsikeresebbnek.) Vesszük a korpuszban egyenként legalább *N*-szer előforduló igék első *N* előfordulását, és kiszámoljuk, hogy egy *f* vonzatkerethez tartozó jeggyel hány ige szerepel egy adott $1 \leq i \leq N$ gyakorisággal. A 3. ábrán a magyar tranzitív keretet jelölő *CAS<ACC>* jegyre vonatkozó statisztika látható. (Részletesebb leírást ld. Serény et al. 2008.)



3. ábra: A tranzitív keretet jelző CAS<ACC> jegy előfordulási valószínűsége a korpuszban szereplő igékkel.

Azt az i_0 gyakoriságot keressük, amelyre igaz, hogy (ebben az esetben) az intranzitív igék többsége i_0 vagy annál kisebb gyakorisággal fordul elő az adott jeggyel, míg a valódi tranzitív igék többsége i_0 vagy annál nagyobb gyakorisággal fordul elő a jeggyel. A megfelelő gyakorisági érték esetén a fenti grafikon bal oldalán egy (ferde) binomiális alakzat jelenik meg. Ebből becsülhetjük meg i_0 értékét, majd az ε hibavalószínűséget.

A hibavalószínűségek ismeretében egy statisztikai modellel döntünk arról, hogy egy ige megjelenthet-e egy adott vonzatkerettel. Három különböző statisztikai modellt alkottunk: binomiális modell, likelihood hányados modell és relativ gyakoriságok.

Binomiális hipotézis próba

Ebben a modellben a nyelvtan kiinduló állapotában minden ige-vonzatkeret párra az áll, hogy egy adott ige *nem* jelenhet meg egy adott vonzatkerettel, és a nyelvtan csak megfelelő pozitív input hatására módosul (konzervatív tanulás).

Az automatikus vonzatkeret-kinyerés feladatának megoldásához először is definiálnunk kell azokat a számszerűsíthető tulajdonságokat, melyek a keresett lexikai információra jellemzőek. A legtöbb módszer az ige és a vonzatjelölt együttes előfordulási statisztikáiból indul ki.

Tehát minden *f* vonzatkerethez hozzárendelünk egy jegykészletet

(4) $f \rightarrow \{c_1^f, c_2^f, \ldots c_n^f\}$

és egy hibavalószínűséget: ε_f, ahol a hibavalószínűség

(5) $\varepsilon_f = P(c_i^f \text{ occurs in } S \mid v \text{ does not take } f)$.

Miután minden keresendő vonzatkerethez rögzítettük jegyek egy halmazát, a következő egyszerű statisztikai modellel döntünk arról, hogy egy ige megjelenhet-e egy adott vonzatkerettel:

(6) $$p_e = P(C(v, f) \geq m \mid v \text{ does not take } f) = \sum_{r=m}^{n} \binom{n}{r} \varepsilon_f^r (1-\varepsilon_f)^{n-r}$$

Veszünk egy *v* igét és egy *f* vonzatkeretet. Nullhipotézisünk, hogy a nyelvtan szerint az ige nem jelenhet meg ezzel a vonzatkerettel. A korpuszban megszámoljuk, hogy az ige hányszor fordul elő összesen (*n*), és hányszor fordul elő a vonzatkerethez tartozó jegyekkel (*C*(*v, f*)). Ha az ige viszonylag sokszor fordul elő a vonzatkerethez tartozó jegyek valamelyikével (p_e kisebb, mint egy előre meghatározott érték), akkor ez arra utal, hogy nullhipotézisünk hibás, a nyelvtan megengedi ezt az ige – vonzatkeret párt. Pontosabban, az ige minden előfordulásakor véletlen kísérlet eredményének tekintjük, hogy egy jegy megjelenik-e vagy nem. A jegy megjelenésének valószínűsége (a nullhipotézis mellett) éppen a jegyhez tartozó hibavalószínűség. A kísérletek eredményei egymástól függetlenek.

Likelihood hányados próba

A gyereknyelvi elemzésekből tudjuk azonban, hogy a vonzatkeretek elsajátítása során túláltalánosításra utaló tanulási mintákat figyelhetünk meg, vagyis az első modell szigorúan konzervatív tanulási algoritmusa valószínűleg nem felel meg a pszicholingvisztikai tényeknek (a modell eredményeit a jelentés 3. fejezetében ismertetjük). Míg az első néhány életévben a gyerek nyelvi produkciójában az ige-vonzatkeret párok száma folyamatosan emelkedik, a helyes argumentumstruktúrák aránya egyes tanulási fázisokban akár csökkenhet is (U-alakú tanulási görbe). Az előbbi mérőszámot a számítógépes nyelvészet "felidézés" (recall) fogalmának, az utóbbit pedig a "pontosság" (precision) fogalmának feleltethetjük meg. Célunk a gyereknyelv és a modell felidézési és pontossági görbéinek egymáshoz való közelítése.

Második modellünkkel olyan statisztikai módszert implementáltunk, amely azt teszteli, hogy egy adott *v* ige megjelenése és egy adott *f* vonzatkerethez tartozó jegy megjelenése egy mondatban független eseményeknek tekinthetők-e, azaz, hogy együttes előfordulásuk mennyire véletlenszerű. Ha a két esemény nem független, *f v* vonzatkeretének tekinthető. A likelihood hányadost a következő egyenlettel definiáljuk:

(7) $$\lambda = l(\frac{k_1 + k_2}{n1 + n_2}, k_1, n_1) + l(\frac{k_1 + k_2}{n1 + n_2}, k_2, n_2) - l(\frac{k1}{n_1}, k_1, n_1) - l(\frac{k_2}{n_2}, k_2, n_2)$$

A függvényben k_1, n_1, k_2, n_2 rendre *v* és *f* jegyének együttes előfordulásának számát, a korpuszban szereplő igék számát, *f* jegyének más igékkel való előfordulásának számát, valamint a *v* igével nem azonos igék számát jelöli. (A modell részletesebb leírását ld. Serény et al. 2008.)

Mivel ez a modell egy adott vonzatkeret más igékkel való előfordulási gyakoriságát érzékenyebben veszi figyelembe, mint az előző modell hibavalószínűségi paramétere, elméletben közelebb áll az emberi nyelvelsajátítás esetében feltételezett általánosító majd a hibás általánosításokat „visszatanuló" tanulási mechanizmushoz.

Relativ gyakoriságok

Harmadik modellünk a Korhonen et al. (2000) által baseline-nak javasolt eljárást valósítja meg. Ez az egyszerű módszer azokat az ege-vonzatkeret párokat fogadja el, ahol a vonzatkerethez tartozó jegyek és az ige együttes előfordulás gyakoriságának az ige előfordulás gyakoriságához viszonyított aránya meghalad egy küszöbértéket. A küszöbértéket empirikus úton határozzuk meg.

3. Eredmények

A három modellt a magyar Webkorpuszon és a Szeged Korpuszon teszteltük. Részleges eredmények láthatók a 2. táblázatban (az eredmények részleteit ld. Serény et al. 2008). Összességében azt állapíthatjuk meg, hogy mindhárom modell teljesítménye jelentősen javul, ha csak a három leggyakoribb vonzatkeretet vesszük figyelembe.

A Brent-féle binomiális módszeren alapuló kísérletet több hibavalószínűségi értékkel is elvégeztük. Az eredmények alapján azt látjuk, hogy ha emeljük a hibavalószínűség értékét, akkor a pontosság megnő, a felidézés értéke viszont csökken. Az F-measure számításakor persze kiegyensúlyozódnak ezek az értékek, de alacsonyabb hibavalószínűségnél összességében jobb teljesítményt kapunk. A 2. fejezetben ismertetett módon előre megbecsült hibavalószínűségi értékekkel számolva rosszabban teljesít a rendszer, mint a fix alacsony hibavalószínűségekkel. A legjobb eredményeket 0.1-es hibavalószínűséggel kaptuk.

A likelihood hányados próba a binomiális módszernél a gépi nyelvfeldolgozás szempontjából kissé gyengébb eredményeket hozott, de a tanulási görbe arra enged következtetni, hogy több tanító adaton (nagyobb korpuszon) a jelenleginél jobban teljesítene. A pszicholingvisztikai párhuzamot tekintve a felidézés magas értéke a pontosság alacsony értékével párosítva a gyereknyelv fejlődésének azt a szakaszát idézi, amikor a kezdeti konzervatív tanulási stratégiát felváltja az általánosító stratégia.

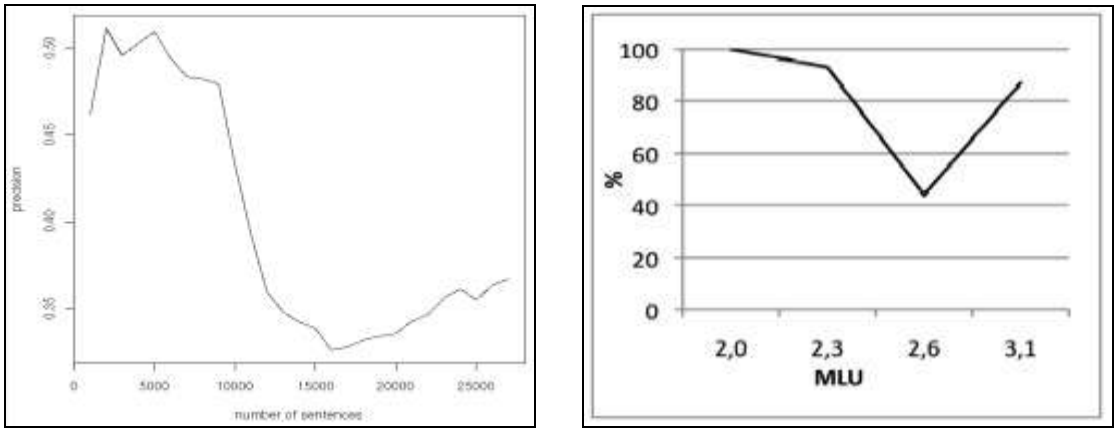
Meglepő módon, a gépi tanulás szempontjából a relativ gyakoriságon alapuló döntés adta a legjobb eredményt.

Módszer	Vonzatkeret	Igék száma	Pontosság	Felidézés	F-measure
Binomiális megbecsült	Trans, dat, ditrans	1000	70%	67%	68%
Binomiális 0.1	Trans, dat, ditrans	200	64%	94%	76%
Relativ gyakoriság	Trans, dat, ditrans	1000	90%	67%	76%
Likelihood próba	Trans, dat, ditrans	1000	25%	79%	39%

2. táblázat: A három modell teljesítménye a három leggyakoribb vonzatkeret elsajátításában.

A pszicholingvisztikai párhuzam szemléltetése érdekében méréseink eredményét grafikusan is ábrázoljuk: a likelihood hányados próba pontossági görbéje (bal

grafikon) hasonló U-alakot mutat, mint a gyerekek tanulási görbéje (jobb grafikon). A tanulási görbe vízszintes tengelyén az idő szerepel (az átlagos mondathosszal jelölve): a kor előrehaladtával a gyerek több input adathoz jut, vagyis tökéletesíteni tudja mentális nyelvtanát és a pontosan használt nyelvtani szerkezetek aránya nő. A likelihood próba eredményének vízszintes tengelyén a korpusz mérete szerepel, ami hasonló funkciót tölt be a gépi tanulás folyamatában. Arra következtetünk, hogy nagyobb korpusz használatával a görbe szára még feljebb kúszna, vagyis még több helyes vonzatkeretet tudna a tanuló algoritmus kivonni a szövegből.



4. ábra: A likelihood hányados próba pontossági görbéje (balra) és három magyar gyerek beszédprodukciójában a *kér* ige helyes vonzatkerettel való használatának aránya (jobbra).

A párhuzam persze távolról sem tökéletes. A 4. ábrán jobbra látható gyereknyelvi U-görbe egy konkrét ige vonzatkeretének fejlődését jeleníti meg, míg a gépi tanulógörbe 1000 ige vonzatkeretének kivonására vonatkozik. A gyereknyelvi korpuszok elemzése során arra az eredményre jutottunk, hogy a jól érzékelhető, szisztematikus vonzatkeret hibák egy-egy igére vagy igecsoportra jellemzőek. A statisztikai gépi tanuláshoz felhasználható korpuszok mérete azonban nem teszi lehetővé, hogy egyes igéket részletesebben tanulmányozzunk.

Hivatkozások

Brent, M R (1993): From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics* 19. 2, pp. 243–262.

Halácsy P., Kornai A., Oravecz Cs. (2007): Hunpos – an open source trigram tagger. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Prague, Czech Republic, pp. 209–212.

Halácsy Péter, Kornai András, Németh László, Rung András, Szakadát István, Trón Viktor (2004): Creating open language resources for Hungarian. In *Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004)*.

Korhonen, A, G. Gorrell and D. McCarthy (2000): Statistical filtering and subcategorization frame acquisition. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*. Hong Kong, pp. 199–206.

Sass Bálint (2006): Extracting Idiomatic Hungarian Verb Frames. In T. Salakoski, F. Ginter, S. Pyysalo, T. Pahikkala (eds.): *Advances in Natural Language Processing*. 5th International Conference on NLP, FinTAL, Turku, Finnország, pp. 303-309.

Serény András, Eszter Simon, Anna Babarczy (2008): Automatic acquisition of Hungarian subcategorization frames. In: Hungaria Fuzzy Association (szerk.) 9th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics (CINTI 2008). Budapest, pp. 443-454.

Trón V., Halácsy P., Rebrus P., Rung A., Vajda P. and Simon E. (2006): Morphdb.hu: Hungarian lexical database and morphological grammar. In *Proceedings of 5th International Conference on Language Resources and Evaluation*. ELRA, pp. 1670–1673.