Magyar igei vonzatkeretek gépi tanulása

Babarczy Anna, Serény András, Simon Eszter

BME GTK Kognitív Tudományi Tanszék,
1111 Budapest, Stoczek utca 2.
e-mail: {babarczy,esimon}@cogsci.bme.hu, andras.sereny@gmail.com

Kivonat A lexikális információ gépi tanulását lehetővé tévő módszerek a számítógépes nyelvészet fontos részterületét alkotják, mert számos természetes nyelvi kétértelműség csak lexikális tudás birtokában oldható fel. Igék esetén ilyen lexikális tulajdonság az is, hogy az ige milyen vonzatkeretekben szerepelhet, azaz milyen kategóriájú bővítményekkel együtt jelenik meg a mondatban. Cikkünkben az igei vonzatkeretek gépi tanulásának más nyelvekre jól működő megközelítéseit, statiszitkai módszereit alkalmazzuk magyar nyelvre. Ezzel párhuzamosan kutatásunknak célja az is, hogy valamilyen módon modellezzük az emberi nyelvelsajátítást, legalábbis a vonzatkeretek elsajátítását; a gépi tanulási görbéket gyereknyelvi adatokból számított tanulási görbékkel vetjük össze.

Kulcsszavak: vonzatkeret-elsajátítás, pszicholingvisztika

1. A lexikális tudás kérdése

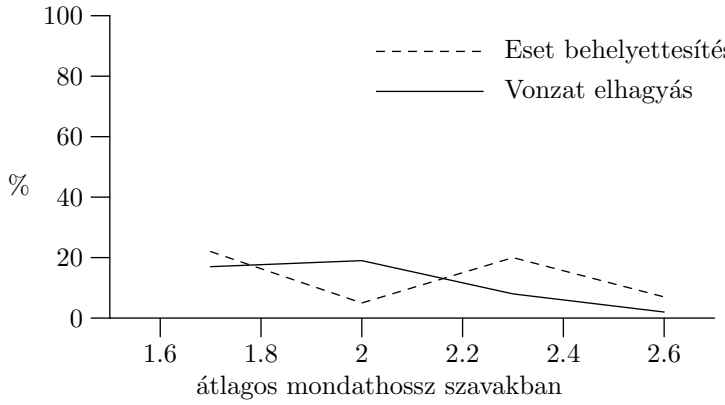
Lexikális tudás elsajátítása alatt a szavak és ezek idioszinkratikus (nem általános elvekből következő) tulajdonságainak elsajátítását értjük, beleértve szemantikai és szintaktikai tulajdonságokat. A predikatív nyelvi elemek – köztük az igék – lexikális tulajdonságai közé tartozik a vonzatszerkezetük, azaz hogy milyen kategóriájú, illetve morfoszintaktikai 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 elemző rendszerek egyik fő problémája. Kutatásunk 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 a mentális lexikon nem az egyes igék vonzatszerkezetének egyenkénti memorizálásával épül, hanem a gyerekek, az input statisztikai tulajdonságait felhasználva, mintákat vonnak ki abból. Ez a tanulás egyes szakaszaiban hibákhoz vezethet. Amint az 1. ábrából kiderül, a gyereknyelvben előforduló vonzatkeretek nem mindig felelnek meg a célnyelvtan által elfogadott vonzatkereteknek.

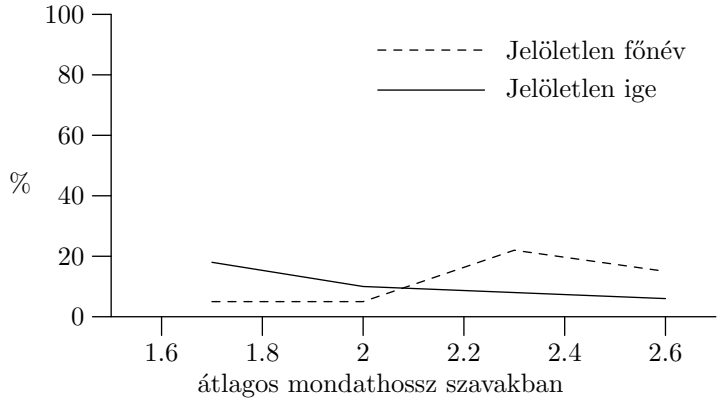


1. ábra. Helytelen nem alanyi esetragok és elhagyott kötelező vonzatok aránya a korai magyar gyereknyelvben. Három gyerek spontán nyelvi produkciójának súlyozatlan átlaga. Korpuszméret: 18 644 szó

A feladatot úgy fogalmazhatjuk meg, hogy ha adott egy F vonzatkeretkészlet és egy V igekészlet, 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 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 vonzatkeretkészlet és egy igekészlet, és a feladata hasonlóképpen az, hogy az igékhez a megfelelő von-

zatkereteket 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 magyarban (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 inflexióelhagyási hiba fordul elő a magyar gyereknyelvben azelőtt is, hogy az átlagos mondathossz elérné a két szót (a jóval gyakoribb morfofonológiai hibákat és ragbehelyettesítéseket itt figyelmen kívül hagyjuk). Feltesszük tehát, hogy a gyerek



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. Korpuszméret: 18 644 szó.

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

2.1. Alapelvek

Kutatásunk fő irányvonala az argumentumstruktúrák elsajátításának számítógépes modellezése volt. A vonzatkeretek gépi tanulására első megközelítésként Brent [1] statisztikai módszerének gazdag morfológiájú nyelvekre adaptált változatát alkalmaztuk. Bár Brent módszere – a számítógépes nyelvészet fejlődési

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

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

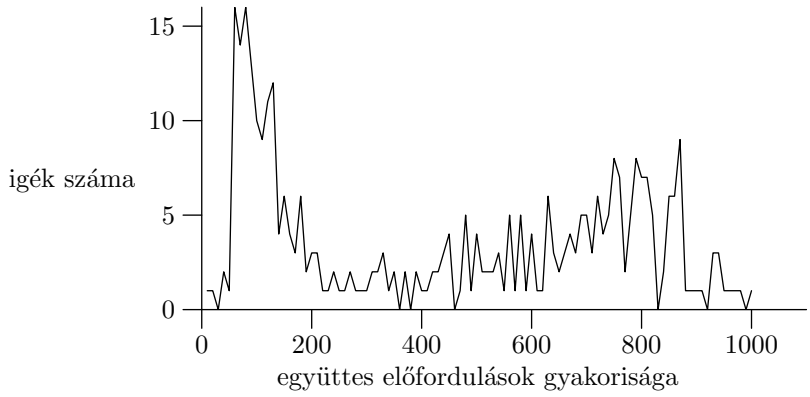
ütemét tekintve – elég réginek nevezhető, magyar vonzatkeretek azonosítására (tudomásunk szerint) ez az első alkalmazása. A magyar nyelvvel foglalkozó munkák közül a miénkhez hasonló tárgyú [6], de ez az idiomatikus, nem kompozicionális, rögzített lemmával előforduló igei szerkezetek kigyűjtését tűzi ki célul. Röviden, Brent eljárásának az a feltételezés az alapja, hogy minden vonzatkerethez tartoznak ún. jegyek. Egy jegy olyan mintázat vagy formai sajátosság, amelynek 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 [10] elemeire illeszkednek: egy jegy illeszkedik egy mondatra, ha a megfelelő reguláris kifejezés illeszkedik a mondathoz tartozó morfológiaiannotáció-sztringre. Az 1. táblázatban egy példát láthatunk. A magyar ditranzitív vonzatkeret például a következő kódnak felel meg:

(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.

2.2. 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. alfejezetben 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 leírásához lásd [8].) Azt az i_0 gyakoriságot keressük, amelyre igaz, hogy (ebben az esetben)



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

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 megjelenhet-e egy adott vonzatkerettel. Három különböző statisztikai modellt próbáltunk ki: binomiális modell, likelihood hányados modell és relatív gyakoriságok.

2.3. Binomiálishipotézis-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

$$f \mapsto \{c_1^f, c_2^f, \ldots c_{n_f}^f\}$$

és egy e_f hibavalószínűséget, ahol a hibavalószínűség

$$e_f = P(c_i^f \text{ megjelenik } S\text{-ben} \mid v\text{-nek } f \text{ nem vonzatkerete})$$

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

e egy adott vonzatkerettel:

$$p_e = P(C(v,f) \geq m \mid v\text{-nek } f \text{ nem vonzatkerete}) = \sum_{r=m}^{n} \binom{n}{r} \varepsilon_f^r (1-\varepsilon_f)^{n-r}.$$

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.

2.4. 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 cikk 3. alfejezetében 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 hányados logaritmusa

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

ahol 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 és a v igével nem azonos igék számát jelöli, valamint $l(q,n,k) = k \log q + (n-k)\log(1-q)$. Ismert, hogy λ eloszlásban tart egy χ^2 eloszláshoz, tehát λ értékeit a χ^2 eloszlás kritikus értékeihez hasonlítva adott szignifikanciájú próbához jutunk. (A modell részletesebb leírását lásd [8].)

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.

2.5. Relatív gyakoriságok

Harmadik modellünk a [5] által baseline-nak javasolt eljárást valósítja meg. Ez az egyszerű módszer azokat az ige–vonzatkeret párokat fogadja el, ahol a vonzatkerethez tartozó jegyek és az ige együttes előfordulási gyakoriságának az ige előfordulási gyakoriságához viszonyított aránya meghalad egy küszöbértéket. A küszöbértéket empirikus úton határozzuk meg.

3. Eredmények

A három modellt a magyar Webkorpusz [4] egy 800 ezer mondatos darabján és a Szeged Korpuszon [2] teszteltük. A Webkorpusz morfológiai annotációját és egyértelműsítését a Hunpos szófaji egyértelműsítővel [3] végeztük. A morfológiai elemzés a KR annotációs nyelvtant használja (ennek részletes leírását lásd [9], [10]). Néhány eredmény látható a 2. táblázatban (az eredmények részleteit lásd [8]). Összességében azt állapíthatjuk meg, hogy mindhárom modell teljesítménye

Módszer	Korpusz	Igék száma	Pontosság	Felidézés	F-mérték
Binomiális	Webkorpusz	1000	70%	67%	68%
Binomiális	Szeged Korpusz	1000	63%	50%	56%
Relatív gyakoriság	Webkorpusz	1000	90%	67%	76%
Likelihood próba	Webkorpusz	1000	25%	79%	39%
Likelihood próba	Szeged Korpusz	1000	35%	56%	43%
Binomiális	Webkorpusz	200	64%	94%	76%
Binomiális	Szeged Korpusz	200	75%	70%	72%

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

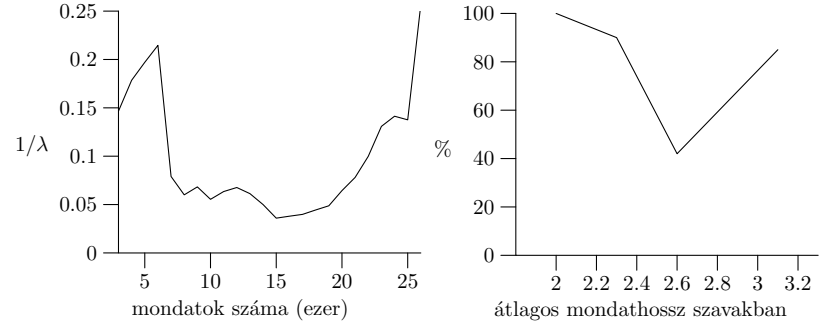
jelentősen javul, ha csak a három leggyakoribb vonzatkeretet vesszük figyelembe. A Brent-féle binomiális módszeren alapuló kísérletet több hibavalószínűségi értékkel is elvégeztük, a táblázatokban elsősorban a 2.2. alfejezetben ismertetett módon előre megbecsült hibavalószínűségi értékekkel számolva kapott értékeket tüntettük fel. Az eredmények alapján azt látjuk, hogy ha emeljük a hibavalószínűség értékét, akkor a pontosság megnő, a felidézés értéke viszont csökken. Az F-mérték számításakor persze kiegyensúlyozódnak ezek az értékek, de alacsonyabb hibavalószínűségnél összességében jobb teljesítményt kapunk. 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	Korpusz	Igék száma	Pontosság	Felidézés	F-mérték
Binomiális	Webkorpusz	100	61%	71%	64%
Binomiális, $\varepsilon = 0,5$	Webkorpusz	100	94%	34%	51%
Relatív gyakoriság	Webkorpusz	100	77%	56%	65%

3. táblázat. A modellek teljesítménye 43 magyar vonzatkeret elsajátításában.

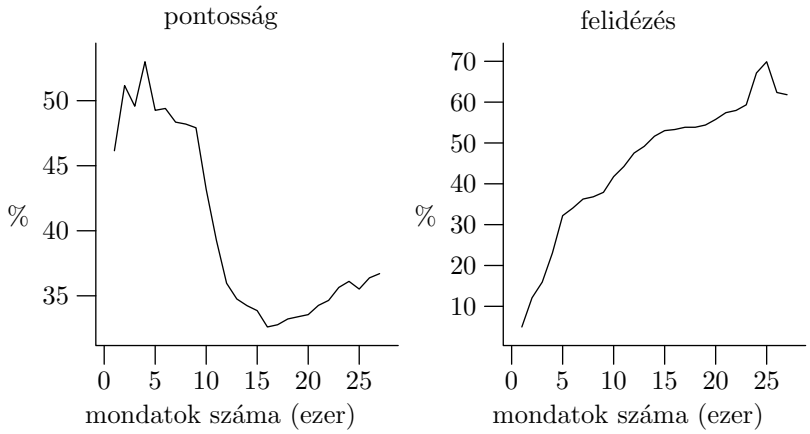
A pszicholingvisztikai párhuzam szemléltetése érdekében méréseink eredményét grafikusan is ábrázoljuk (4. ábra). A likelihood statisztika $1/\lambda$ reciproka jó mérőszáma annak, hogy a modell egy adott ige–vonzatkeret párt „mennyire" gondol helyesnek. Ez a görbe (bal grafikon) hasonló U-alakot mutat, 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 bemeneti 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. A gyereknyelvi korpuszok elemzése során arra az



4. ábra. A likelihood statisztika görbéje a *kér – kér valamiből* ige–vonzatkeret párra a Szeged Korpuszon (balra) és három magyar gyerek beszédprodukciójában a *kér* ige helyes vonzatkerettel való használatának aránya (jobbra).

eredményre jutottunk, hogy a jól érzékelhető, szisztematikus vonzatkerethibák

egy-egy igére vagy igecsoportra jellemzőek. Az 5. ábrán a likelihood hányados próba pontosságát és felidézését láthatjuk. A pontosság a kezdeti csökkenés után növekedésnek indul. Arra következtetünk, hogy nagyobb korpusz használatával a görbe szára még feljebb kúszna, vagyis még több helyes vonzatkeretet tudna a tanuló algoritmus kivonni a szövegből.



5. ábra. A likelihood hányados próba pontossága és felidézése a Szeged Korpuszon

Hivatkozások

1. Brent, M. R.: From grammar to lexicon: Unsupervised learning of lexical syntax. Computational Linguistics 19, 2(1993) 243–262
2. Csendes, D., Csirik, J., Gyimóthy, T.: The Szeged Corpus: A POS tagged and Syntactically Annotated Hungarian Natural Language Corpus. In: Proceedings of TSD 2004. Brno, vol. 3206 (2004)
3. Halácsy, P., Kornai, A., Oravecz, Cs.: 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 (2007) 209–212
4. Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., Trón, V.: Creating open language resources for Hungarian. In: Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004) (2004)
5. Korhonen, A, G. Gorrell, McCarthy, D.: 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 (2000) 199–206
6. Sass, B.: Extracting Idiomatic Hungarian Verb Frames. In: Salakoski, T., Ginter, F., Pyysalo, S., Pahikkala, T. (eds.): Advances in Natural Language Processing. 5th International Conference on NLP, FinTAL, Turku, Finnország (2006) 303–309

7. Schulte im Walde, S.: The induction of verb frames and verb classes from corpora. In: Lüdeling, A., Kytö, M. (eds.): Corpus Linguistics. An International Handbook. Berlin, Mouton de Gruyter (2008)
8. Serény, A., Simon, E., Babarczy, A.: Automatic acquisition of Hungarian subcategorization frames. In: Hungarian Fuzzy Association 9th International Symposium of Hungarian Researchers on Computational Intelligence and Informatics (CINTI 2008), Budapest (2008) 443–454
9. Trón, V., Halácsy, P., Rebrus, P., Rung, A., Vajda, P., Simon, E.: Morphdb.hu: Hungarian lexical database and morphological grammar. In: Proceedings of 5th International Conference on Language Resources and Evaluation. ELRA (2006) 1670–1673
10. Kornai A., Rebrus P., Vajda P., Halácsy P., Rung A., Trón V.: Általános célú morfológiai elemző kimeneti formalizmusa. In: Alexin Z., Csendes D. (szerk.): II. Magyar Számítógépes Nyelvészeti Konferencia. SZTE Informatikai Tanszékcsoport, Szeged (2004) 172–176
11. Zeman, D., Sarkar, A.: Automatic extraction of subcategorization frames for Czech. In: Proceedings of the International Conference on Computational Linguistics (COLING '00)(2000) 691–697