Természetes nyelvi korpusz vizsgálata egyeztetés-csoport módszerrel

Drienkó László

dri@t-online.hu

Kivonat: Korábbi munkáimból, [1, 2, 3], kiindulva egy olyan disztribúciós módszert szeretnék bemutatni, amely alkalmas lehet természetes (akár nem természetes) nyelvi szekvenciákat tartalmazó adatbázisokban rejlő, az adott nyelvre jellemző bizonyos disztribúciós szabályosságok detektálására, illetve új szekvenciáknak ezen szabályosságokon alapuló feldolgozására, a szabályosságokat megtestesítő szekvencia csoportokra való "leképzésére". Az alapvető fogalmak felvázolása, illetve a korábbi eredmények rövid ismertetése után néhány megjegyzés következik a módszer alkalmazhatóságával kapcsolatban, melyek nyomán az egyeztetés-csoportok által képviselt kutatás spektruma szélesedhet.

1 Bevezetés

Az egyeztetés-csoport módszer lényege, hogy egy „training" szekvencia-halmaz elemeiből csoportokat hozunk létre, és megvizsgáljuk, milyen mértékben képezhetők le egy tetszőleges új halmaz szekvenciái ezekre a csoportokra. A csoportosítás minimális különbségen alapul, azaz minden szekvenciához megkeressük azokat a szekvenciákat, amelyek csak egyetlen elemben térnek el tőle. Az „egyeztetés"- csoport elnevezés azon megfigyelésen/feltételezésen alapul, hogy amennyiben egy mondatban egy tetszőleges szót egy ugyanolyan „lexikai" kategóriájú szóval helyettesítünk és az így kapott új mondat nyelvtanilag helyes, akkor az eredeti mondat egyeztetés-viszonyai meg kell hogy őrződjenek, mivel a behelyettesített szónak rendelkeznie kell az eredeti mondat által megkövetelt egyeztetés-jegyekkel.

Az adott csoportokat táblázatos formában reprezentáljuk, ami a nyelvtani „következtetés", azaz az új mondatok feldolgozásának az alapja lesz. Például az (1)-ben megadott egyeztetés-csoport (2)-beli táblázatos formája lehetővé teszi (3) új mondatainak feldolgozását, azaz (1)-re való „leképzését".

(1)

Adam hates football
Adam hates **basketball**
Eve hates football
Adam **dislikes** football
Charles hates football

(2)

Adam	hates	football
Eve	dislikes	basketball
Charles		

(3)

Eve hates basketball	Eve dislikes football
Charles hates basketball	Charles dislikes football
Eve dislikes basketball	Charles dislikes basketball
Adam dislikes basketball	

2 Korábbi eredmények

2.1 Egyeztetés-csoport analízis

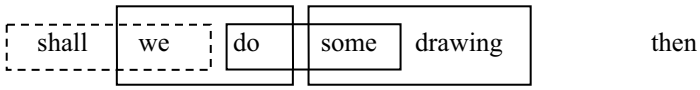
A módszert CHILDES [4] angol [5], magyar [6] és spanyol [7] gyermeknyelvi adatokra alkalmaztam. Mindegyik nyelv esetében az egyes felvételek idejét az adott anya-gyermek nyelv egy bizonyos fejlődési állapotának feleltettem meg és az addig elhangzott összes 2-5 szavas kijelentésből formáltam egyeztetés-csoportokat. Ezután megvizsgáltam, hogy a közvetlenül következő felvétel 2-5 szavas kijelentéseinek hány százaléka képezhető le az adott állapothoz tartozó egyeztetés-csoportokra. Azt találtam, hogy mindegyik fejlődési állapotban a kijelentések bizonyos hányada leképezhető a már meglévő csoportokra – enyhe növekedés is megfigyelhető. A leképzési értékek az angol esetben voltak maximálisak: egy esetben a soron következő felvétel 41% volt megfeleltethető az előzőekből nyert egyeztetés-csoportoknak. Az új kijelentések leképzési maximuma 10,3% volt.

2.2 Egyeztetés-csoport lefedhetőség

Következő lépésként [8,9] megvizsgáltam, hogyan lehet hosszabb kijelentéseket „lefedni” 2-5 szavas, egyeztetési csoportokra leképezhető szekvenciákkal. Például a

'shall we do some drawing then' mondat lefedhető a *'shall we'*, *'we do', 'do some' 'some drawing'* szekvenciákkal. Vö. (4).

(4)



Egy kijelentés lefedettség értékének kiszámítási módját (5)-ben vázoljuk.

(5)

1	2	3	4	5	6
shall	we	do	some	drawing	then
1	1	1	1	1	0

Lefedettség (Coverage): (5 szó a 6-ból): 5/6=83%

Az angol CHILDES adatokra 78% átlagos lefedettséget kaptam [8], a magyar nyelvi esetben 42%-ot [9]. A (4)-beli szaggatott vonal azt jelzi, hogy a 'shall we' szekvencia már konkrétan szerepelt a „training" halmazban, azaz nem új. Viszont megvizsgálhatjuk, hogy kifejezetten az új szekvenciák milyen mértékben járulnak hozzá a lefedettséghez. Ekkor pl. az (5)-beli érték 4/6= 66% lesz, mivel az 1. pozícióhoz 0 rendelődik. Az új szekvenciák az angol adatokra 49%-os, a magyar adatokra 29%-os átlagos lefedettséget eredményeztek.

Elvi síkon, a fent vázolt kísérletek kiindulópontjául szolgálhatnak egy olyan kétszintű nyelvi/nyelvtanulási modellnek, ahol egy alapvetőbb, elsődleges kognitív szint felel az egyszerűbb, kevésbé komplex megnyilvánulások feldolgozásáért – esetünkben ezt a szintet az egyeztetés-csoportok képviselik –, ugyanakkor a komplexebb struktúrák létrehozása, azaz az egyszerűbb megnyilvánulások egymáshoz rendelése, integrációja egy magasabb kognitív szinten történik. Hosszabb mondatoknak rövidebb fragmentumokkal való lefedhetőségét vizsgáló kísérleteink e második szintet próbálták vizsgálni.

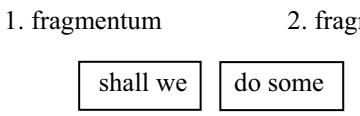
3 A módszer alkalmazhatóságával kapcsolatos megjegyzések

3.1 Fragmentum kombinációk

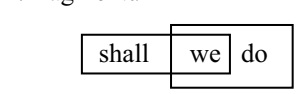
Mind nyelvelméleti, mind számítástechnikai szempontból fontos kérdés lehet a lefedhetőséget illetően, hogy milyen fragmentum kombinációk eredményezhetnek nyelvileg helyes megnyilvánulásokat. Ahogyan az (4)-ből is látszik, algoritmusunk alapjában véve kétféle fragmentum konfigurációt ismert fel, mivel feltételezte a fragmen-

tumok folytonosságát, vagyis azt, hogy bármely fragmentum bármely két eleme (szava) között nincs más fragmentumhoz tatozó elem. Ennek legkézenfekvőbb esetét mutatja (6), ahol a fragmentumok jól elhatároltan követik egymást. A folytonosság feltételezése persze nem zárja ki, hogy két fragmentum közé más elemek kerüljenek, illetve hogy bizonyos szélső elemek mindkét fragmentumhoz tartozzanak, mint például a '*we'* szó (7)-ben.

(6)



(7)

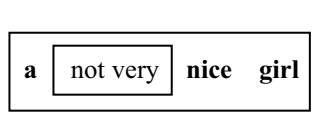


Elképzelhető azonban, hogy a fragmentum-folytonosság feltételezésének feladása nagyobb lefedettség értékekhez vezethetne, mivel összetettebb nyelvi szerkezetek is közvetlenül elérhetőek lennének. (8) például azt mutatja, hogyan ágyazhatók be egymásba fragmentumok: (8a)-ban nincs közös elem, (8b)-ben viszont a *nice* szó mindkét fragmentumhoz hozzátartozik. (8c) azt vázolja, hogyan fedhető le a klasszikus *The rat the cat the dog bit chased ate the cheese* szerkezet három fragmentummal. (9) példái egyfajta keresztfüggőség (cross serial dependency) hatást érzékeltetnek kettő (vö. 9a), illetve három (vö. 9b) fragmentummal.

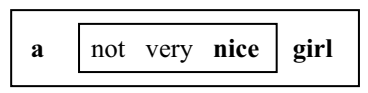
(10)-ben a beágyazás és a keresztfüggőség lehetséges kombinálására mutatunk példát: az első és második, illetve az első és harmadik fragmentum viszonylatában keresztfüggőséget látunk, ugyanakkor a harmadik fragmentum beágyazódik a másodikba.

(8)

a)



b)



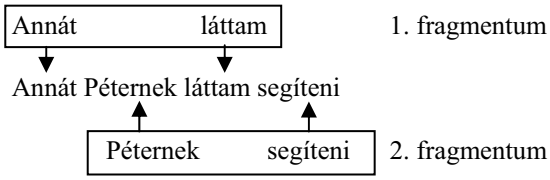
c)

1. fragmentum: **The rat ate the cheese** 2. fragmentum: *The cat chased*
3. fragmentum: The dog bit

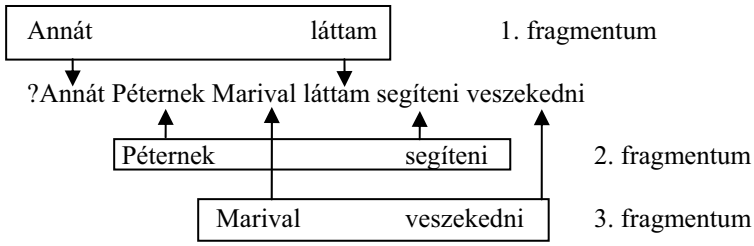
(**The rat** (*The cat* (The dog bit) *chased*) **ate the cheese**)

(9)

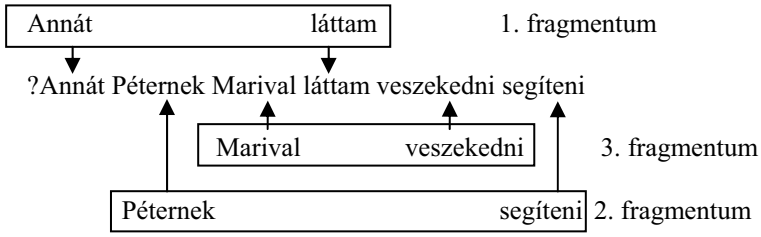
a)



b)



(10)



A fragmentum-folytonosság feltételezés feladása a jelenlegi algoritmus módosítását igényelné, ami viszont a számítástechnikai erőforrások bővítését tenné szükségessé.

3.2 Az egyeztetés-csoportok mint kognitív nyelvi struktúrák

A bemeneti adathalmaz strukturálásának eredményeképpen létrejövő egyeztetés-csoportok önmagukban is hordozhatnak releváns pszicholingvisztikai, illetve kognitív nyelvészeti információt az adott nyelv felépítését illetően. Az egyes csoportok mini-

málisan különböző szekvenciái alapjául szolgálhatnak olyan kategorizálási mechanizmusoknak, amelyek felelősek lehetnek a nyelv különböző szintű – lexikai/szintatikai, szemantikai, fonetikai, stb. – kategóriáinak kialakulásáért, továbbá a különböző nyelvi szintekhez tartozó jegyegyeztetési folyamatokért. (11a) mondataiban például az utolsó szó pozíciót igék töltik be, csakúgy mint (11b)-ben, ahol viszont szembeötlőbb az alany-ige egyeztetés – egyes szám első személy. A (11c)-beli főneveket szemantikailag egyfajta helymegjelölő funkció kapcsolja össze.

(11)

a)

you can't **reach** you can **reach** you can't **remember** you can't **know**
I can't **reach** you can't **see** you can't **play**

b)

*én nem én én nem **látom** én nem **játszok** én nem **tudom**
én nem **kéjek** én nem **szejetem** én nem **vagyok** én nem **tudok**
én nem **láttam**

c)

to the **shops** to the **hospital** to the **pub** to the **garden**
to the **seaside** to the **car** to the **farmyard**

Végül megemlítjük, hogy az egyeztetés-csoport módszer elvileg bármely olyan szekvencia halmaz esetén alkalmazható, ahol feltételezhető, hogy az egyes szekvenciák jólformáltságáért valamilyen mögöttes „szekvencia gyártó” mechanizmus felel.

Hivatkozások

1. Drienkó, L.: Agreement groups analysis of mother-child discourse. Talk presented at the 4th UK Cognitive Linguistics Conference, King's College London, UK (2012). To appear in Selected Papers from UK-CLA Meetings. Vol. 2.
2. Drienkó, L.: A linguistic agreement mapping-system model: agreement relations for linguistic processing. LAP-Lambert Academic Publishing (2012)
3. Drienkó, L.: Distributional cues for language acquisition: a cross-linguistic agreement groups analysis. Poster presentation for the 11th International Symposium of Psycholinguistics, Tenerife, Spain (2013)
4. MacWhinney, B.: The CHILDES Project: Tools for analyzing talk. 3rd Edition. Vol. 2: The Database. Mahwah, NJ: Lawrence Erlbaum Associates (2000)
5. Theakston AL, Lieven EV, Pine JM, Rowland CF.: The role of performance limitations in the acquisition of verb-argument structure: an alternative account. J. Child Lang. 28(1): (2001) 127–52
6. Réger, Z.: The functions of imitation in child language. Applied Psycholinguistics 7. (1986) 323–352

7. Montes, R. G.: Achieving understanding: Repair mechanisms in mother–child conversations. Unpublished doctoral dissertation, Georgetown University (1992)
8. Drienkó, L.: Agreement groups coverage of mother-child language. Talk presented at the Child Language Seminar, Manchester, UK (2013)
9. Drienkó, L.: Agreement groups coverage of Hungarian mother-child language. Poster presentation for the 11th International Conference on the Structure of Hungarian. Piliscsaba, Hungary (2013)