Néhány nyelvstatisztikai módszerrel végzett elemzés összehasonlítása

Bujdosó Iván

ELTE BTK Alkalmazott Nyelvészeti Tanszék
bujdosxo@yahoo.com

1. Bevezetés

Az elemzés a szóstatisztikai elemzés körére korlátozódik. Az elemzési módszerek:

- egyszerű összeszámlálás és elemi matematikai műveletek
- normál eloszlási görbe, átlag és szórás számítása
- hatvány törvény alkalmazása
- mesterséges neurális hálózatok alkalmazása

A vizsgált nyelvek: az Európai Unió összes hivatalos nyelve és az eszperantó. A vizsgált nyelvek a kapott számszerű értékek szerint sorrendbe rakhatók. A vizsgálat végeredményének megelőlegezésével a nyelvek sorrendje:

1. finn; 2. észt; 3. magyar; 4. litván; 5. lett; 6. szlovák; 7. cseh; 8. lengyel; 9. szlovén; 10. máltai; 11. eszperantó; 12. görög; 13. dán; 14. svéd; 15. német; 16. olasz; 17. portugál; 18. spanyol; 19. francia; 20. holland; 21. angol

A vizsgált szövegek:

- Szerződés európai alkotmány létrehozásáról I-II. rész. A nyelvenként mintegy 35 oldalas anyag az összes felsorolt nyelven megtalálható az interneten.
- 283 regény angolul, eszperantóul, franciául, németül, olaszul és spanyolul; az internetről letöltve

Hipotézis: A nyelvek jellemezhetők egy, a vizsgálati módszertől függő számértékkel, és ezek alapján a nyelvek vizsgálati módszerenként sorba rendezhetők. Az így kapott sorok majdnem teljesen azonosak. Ezért az egyes vizsgálati módszerek a nyelv jellemzésére alkalmasak.

A vizsgálati módszerek közös jellemzője a szószintű vizsgálat. A szövegekből eltávolítottam a tagolásra szolgáló jeleket.

Egy szöveg szavai az alábbiakkal jellemezhetők: szóhossz, előfordulási gyakoriság, majd ezek alapján a szavak rangsora. Képzett értékek: a csak egyszer előforduló szavak száma (hapaxok), szókettősök, szóhármasok, szótávolság (ugyanazon szó következő előfordulása), Zipf egyenes jellemzői (meredekség, konstans, a regressziós együttható: R2), stb.

2. A négy különböző elvű vizsgálat:

Az első vizsgálatban az egyes nyelvnél a hapaxok számát arányítottam az összes szó számához. A két szélső értéket a finn (60%), illetőleg az angol (44%) adta.

A második vizsgálatban mind a 21 nyelvnek a szövegben előforduló szavát egy halmazként kezelve, megállapítottam egy "európai szóhosszúsági átlagot". Ebből kivontam az egyes nyelvekre jellemző értékeket (az egy betűs hosszúságú európai átlagból a kérdéses nyelv egy betűs szavainak a számát, ugyanígy a két-, a három-, stb-betűs szavak számát). Nyelvenként átlagot képeztem.

A harmadik vizsgálatnál minden nyelvnél megállapítottam a Zipf-görbe meredekségét. A két szélső értéket a finn (-0,76) és az angol (-1,11) adta.

A negyedik vizsgálatot egy amerikai egyetemen végezte el Manaris és munkatársai. Ők a mesterséges neurális hálók elméletét használták, a vizsgált 283 könyv szövegének terjedelme több nagyságrenddel nagyobb az alkotmánytervezet szövegénél. A nyelvek sorrendje azonban ugyanaz lett, mint az általam megállapított sorrend. Ezen túlmenően a kapott számszerű értékek és az általam végzett vizsgálat értékei jó egyezést mutattak, spanyol nyelv figyelembe vétele nélkül R2 = 0,98, spanyol nyelv figyelembe vételével R2 = 0,73.

A négyféle vizsgálat eredményeit láthatjuk az alábbi táblázatban a kapott számszerű jellemzőknek megfelelően. A táblázatban az egyes nyelveket az előző felsorolás számértékei jelentik, azaz a magyar nyelv az első vizsgálat szerint a 3., a második vizsgálat szerint a 9., a harmadik vizsgálat szerint a 3. helyen áll, a negyedik vizsgálatban nem szerepelt. A táblázatban a magyar nyelv kódját félkövérrel, az eszperantóét dőlt karakterrel szerepeltetem.

1	2	**3**	4	5	12	7	10	6	*11*	8	16	5	13	17	14	19	15	18	20	21
1	2	5	4	7	9	8	6	**3**	10	13	14	15	16	12	*11*	17	19	21	20	18
1	2	**3**	4	5	6	7	8	9	10	*11*	12	13	14	15	16	17	18	19	20	21
~	~	~	~	~	~	~	~	~	~	*11*	~	~	~	15	16	~	18	19	~	21

(Ahol ~ azt jelenti, hogy nincs vizsgált anyag erre a nyelvre, a vizsgálat Manaris szerint).

3. A fenti vizsgálatok eredményei:

1. Már egészen kis méretű korpuszból is következtethetünk arra, hogy a kérdéses szöveg milyen nyelven íródott.
2. A teljesen azonos tartalmú szövegek Zipf-együtthatóit sorrendbe állítva a nyelvrokonságról szóló ismereteinkkel összhangban levő képet adnak (a finnugor, a szláv, a germán, az újlatin nyelvek egymás mellett vannak, valamint a balti nyelvek a szláv nyelvek mellett). Ez az eredmény igazolja a kezdetben felállított hipotézisünket, azaz a kapott eredmény nem lehet a véletlen műve. Az eszperantó esetében pedig számszerű igazolását adja

Pennacchietti professzor 1981-ben kifejtett véleményének[84], amelyben az eszperantót tipológiailag a szláv és a germán nyelvek közé teszi, ellentétben azokkal a vélekedésekkel, amelyek agglutináló vagy izoláló jellegzetességeit hangsúlyozták.

Bibliográfia

1. Gledhill, C. 1998. The Grammar of Esperanto. A corpus-based description. München: Lincom Europa.
2. Manaris et al. 2006: Investigating esperanto's statistical proportions relative to other languages using neural networks and Zipf's law. Proceedings of the 2006 IASTED International Conference on ARTIFICIAL INTELLIGENCE AND APPLICATIONS (AIA 2006), February 13-16, Innsbruck, Austria.
3. Pennacchietti, F. 1981: Ne-hindeŭropaj trajtoj de la internacia lingvo, in: Sprachkybernetik, 1981, Paderborn, p. 95

[84] "La interna kohereco de Esperanto klariĝas do per tio, ke ĝi kapablas harmoniigi la postulojn de struktura simpleco, necesajn por vasta internacia uzo, kun la konservado de preciza *tipologia* stampo, nome tiu de la *ĝermanaj* kaj *slavaj* lingvoj de centra Eŭropo."