Korpuszalapú entrópiamértékek gating- és lexikai döntési kísérletekben

Fazekas Judit[1], Németh Kornél[1], Pléh Csaba[1], Varga Dániel[2]

[1] BME Kognitív Tudományi Tanszék, Budapest, Egry József utca 1.
e-mail: {jfazekas,knemeth,pleh}@cogsci.bme.hu
[2] BME MOKK, Budapest, Egry József utca 1.
e-mail: daniel@mokk.bme.hu

Nagyméretű gyakorisági szótár birtokában lehetőségünk nyílik információelméleti mértékeket definiálni, amelyek olyan kérdéseket formalizálnak, mint például hogy egy adott szó-prefix a korpuszon belül milyen mértékben korlátozza a szó lehetséges befejezéseinek halmazát.

Cikkünkben ezen mértékek felhasználásával megkíséreljük, hogy összefüggést tárjunk fel az emberi morfológiai feldolgozás és szófelismerés teljesítménye, valamit a szóalakok információelméleti struktúrája között.

Cikkünk bővített változatában három olyan kísérlet eredményeit mutatjuk meg, melyek a fenti kérdéseket járják körül szisztematikus módon.

Az első két, gating feladaton [5] alapuló vizsgálat anyagát 60 darab kétszótagú főnév képezte. A 30 gyakori és a 30 ritka szó közül 15-15 korai egyediségi ponttal rendelkezett (*japán*), 15-15 pedig későivel (*cinke*). A varianciaanalízis egyedül a gyakoriságról mutatta ki, hogy szignifikáns hatása van a felismerés hatásfokára.

A második vizsgálatban bevezettünk egy megszorítást, a szófelismerést befolyásoló top-down hatások vizsgálatának céljából. A résztvevők fele a következő instrukciót kapta: "Csak kétszótagú főneveket fog hallani toldalékok nélkül.", a többi kísérleti személy nem kapott semmilyen információt. Mind a gyakoriság, mind pedig a megszorítások hatása kimutatható volt. Az egyediségi pontok hatása csak a gyakori szavaknál volt egyértelmű.

A mérési adatok birtokában az egyértelműségi pont fogalmának korpuszalapú finomítása céljából a Magyar Webkorpuszra épülő morfológiailag elemzett Szószablya Gyakorisági Szótárhoz [3] fordultunk, és a szótár prefix-fájának információelméleti analízisét végeztük el. Ennek során entrópiamértéket vezettünk be a szóalakok prefixein, az alábbi módon: A gyakorisági szótár a magyar nyelv szóalakjain értelmezett valószínűségeloszlást definiál. Egy szó-prefix entrópiáját ezután úgy definiáltuk, mint e valószínűségeloszlásnak a feltételes entrópiáját azon feltétel mellett, hogy a megfigyelt szó az adott prefixszel kezdődik. A feltételes entrópia tehát a fennmaradó bizonytalanságunk mértéke az adott szóról, miután a prefixét a tudomásunkra hozták. Intuitíve, a mérték azt számszerűsíti, hogy mennyire változatos módon fejeződhet be az adott prefix a korpuszunkban.

Megemlítjük, hogy Antal László [2] már 1964-ben felvetette azt a hipotézist, hogy a morfológiailag összetett szavak morfémahatárai statisztikai értelemben összefüggésbe hozhatók azon pontokkal, ahol az így definiált entrópiamérték zu-

han. A Szószablya Gyakorisági Szótáron végzett méréseink igazolták ezt a hipotézist.

Egy adott kapuhoz az ott felvett mérési pontokat három osztályba soroltuk, aszerint, hogy 1. éppen abban a pontban történt meg a felismerés, 2. éppen a következő pontban történt meg a felismerés, illetve 3. egyéb esetek. Azt tapasztaltuk, hogy valamely kaput rögzítve, a prefixek entrópiamértéke szignifikáns mértékben eltér az 1. és 2. kategóriájú adatpontok között, vagyis a felismerést még a kapura kontrollálva is entrópiacsökkenés előzi meg. Ez a jelenség még akkor is fennáll, ha a gyakoriságra és az egyediségi pont helyére mint kétértékű változókra kontrollálunk. Mi ezt a megfigyelést úgy értelmezzük, mint amely demonstrálja, hogy az entrópia szándékainknak megfelelően az egyediségi pont naív fogalmának kvantitatív finomítása. Ez az eredmény összhangban van Moscoso, Kostic és Baayen [4] modelljével.

Nemcsak az entrópia, hanem az entrópia szomszédos kapuk közötti megváltozása is mutatta a fenti jelenséget, annak ellenére, hogy ez egy erősen nemmonoton viselkedést mutató függvény.

Egy következő kísérletünk Pléh és Juhász [6] szófelismerésre vonatkozó vizsgálatainak folytatása volt. Itt rontott szavak azonosítása volt a kísérleti személyek feladata. A szavak egyes vagy többes számúak voltak, tőalakban, vagy a -nak, -ban, -ra ragokkal. A rontás a szótő, a jel, illetve az esetrag valamelyikében történt, és típusukban lehettek magánhangzó-harmónia hibák, vagy a szótőben történő fonémarontások.

A gyakoriságnak és a rontás típusának egyaránt szignifikáns hatása volt az azonosítás pontosságára. A gyakoribb szavakat gyorsabban kategorizálták a kísérleti személyek, de alacsonyabb pontossággal. Erős korreláció volt a rontás pozíciója és a sikeres visszautasítások aránya között; a későbbi rontások gyorsabb és pontosabb visszautasításhoz vezettek.

Gyakorisági szótárunk segítségével korpuszalapú vizsgálatnak is alávetettük ezen mérések kimeneteit. Hipotézisünk az volt, hogy könnyebben felismerhetőek azok a rontások, melyek szokatlan fonéma n-gram kombinációkhoz vezetnek. A hipotézis formalizálásához meghatároztuk a fonéma trigramok gyakoriságait a korpuszunkban, majd metrikánkat úgy definiáltuk, mint a rontás fonéma trigram környezetének gyakorisága arányítva az eredeti, rontatlan fonéma trigram környezet gyakoriságával. Hipotézisünknek megfelelően a sikeres visszautasítás valószínűsége és sebessége egyaránt erős korrelációban volt az így definiált rontásitrigram-metrikával.

Hivatkozások

1. Aitchison, J.: Words in the mind. London, Blackwell (1987)
2. Antal, L.: A formális nyelvi elemzés, Budapest, Gondolat (1964)
3. Kornai, A., Halácsy. P., Nagy. V., Oravecz, Cs., Trón, V., Varga, D.: Web-based frequency dictionaries for medium density languages. In: Proceedings of the EACL 2006 Workshop on Web as a Corpus (2006)

4. Moscoso, F., Kostic, A., and Baayen, R. H.: Putting the bits together: an information theoretical perspective on morphological processing. Cognition, 94, pp. 1-18 (2004)
5. Grosjean, F.: Spoken word recognition processes and the gating paradigm. In: Attention, Perception, & Psychophysics, Springer (1980)
6. Pléh, Cs., Juhász, L. Processing of multimorphemic words in Hungarian. Acta Linguistica Hungarica, 43, pp. 211-230. (1995)