Tóth Ágoston: Magyar nyelvű szóhasonlósági mérések humán és disztribúciós szemantikai kísérletben
Argumentum 9 (2013), 301-310
Debreceni Egyetemi Kiadó

Tóth Ágoston

Az ember, a korpusz és a számítógép*

Magyar nyelvű szóhasonlósági mérések humán és disztribúciós szemantikai kísérletben

Abstract

The paper reports on the results of two word similarity experiments. The first experiment is a subjective human test: similarity values for 31 pairs of Hungarian words have been collected from 28 subjects. The test method comes from Rubenstein & Goodenough (1965) and it reflects the intuition that word similarity is a continuum from clear cases of synonymy to the complete lack of apparent similarity. The Hungarian results correlate very well with the data collected by Rubenstein and Goodenough (Spearman $r=0,959$, $p<0,01$) and also with the English replica experiments (Miller & Charles 1991 and Resnik 1995). In the second experiment presented here, a computer program collected similarity data for the same words, based on the context in which they typically occur. The correlation between the subjective and the corpus-based data series is $r=0,591$ ($p<0,01$).
Keywords: word similarity, distributional semantics, vector spaces, computational linguistics

1 Bevezetés

A szóhasonlóság automatikus kiszámításának két elterjedten használt módszere létezik: lexikai adatbázisok és ontológiák felhasználása, valamint statisztikai adatok kinyerése nagyméretű korpuszokból.[1]

A *lexikai adatbázisok* használatának előnye, hogy szakértők által előkészített, megbízható és rendszerezett információval dolgozunk. Ezekből az adatokból közvetlenül kiszámítható a kérdéses szavak (fogalmak) távolsága, például alá- és fölérendeltségi hierarchiákban megkeresve a fogalmakat, és az őket összekötő útvonal mentén megszámolva az éleket, esetleg ennek a módszernek valamilyen továbbfejlesztett, súlyozott változatát használva (pl. Resnik 1995). Ezeket a megoldásokat természetesen érinti a jelentésfelsoroló lexikonok összes potenciális hátránya, de leginkább azért nehéz őket a gyakorlatban alkalmazni, mert a megfelelő fogalmat a feldolgozás első lépéseként be kell azonosítani az ontológiában, ami leginkább manuálisan, vagy a meglehetősen pontatlan automatikus jelentés-egyértelműsítő rendszerek (v.ö. Tóth 2011) használatával történhet.

* A kutatás a TÁMOP 4.2.4.A/2-11-1-2012-0001 azonosító számú *Nemzeti Kiválóság Program – Hazai hallgatói, illetve kutatói személyi támogatást biztosító rendszer kidolgozása és működtetése konvergencia program* című kiemelt projekt keretében zajlott. A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg. A cikk az OTKA K 72983 számú projektjének támogatásával jött létre.

1 Ezúton köszönöm Hollósy Bélának, hogy bevezetett a korpusznyelvészet, a lexikai adatbázisok és általában a számítógépes nyelvészet világába.

Ezzel a hátránnyal nem rendelkeznek a *nagyméretű korpuszokkal* dolgozó, a szavakat a környezetükben előforduló további szavak segítségével jellemző, majd ezen tulajdonságukat összehasonlító („vektorteres", disztribúciós szemantikán alapuló) algoritmusok. Ebben az esetben azt vizsgáljuk, hogy adott szavak az adott pozícióban mennyire szokták egymást helyettesíteni (korpuszból nyert evidencia alapján). Mivel ebben a megközelítésben szóalakokkal vagy lemmákkal dolgozunk, ezért a megfelelő fogalom megkeresése nem az algoritmus működésének az előfeltétele. További előny, hogy az algoritmusok ezen osztálya „csupán" egy nagyméretű, sokszor annotációt sem tartalmazó korpuszt igényel, lexikai adatbázisokra nem támaszkodik. Ez azért is fontos, mert így távolabbi célként kitűzhető a *szavak jelentésének automatikus elsajátítása*, ami mind nyelvészeti, mind mesterséges intelligencia alkalmazásokban fontos szerepet kaphat.

Bármelyik automatikus szóhasonlósági mérést szeretnénk alkalmazni, felmerül annak a kérdése, hogy az így kapott eredményeinket hogyan értékeljük ki, mihez hasonlítsuk. A disztribúciós szemantika szakirodalmában a következő eljárások a leggyakoribbak (Bullinaria & Levy 2007 alapján):

- távolság összehasonlítása: feleletválasztás, melyben adott célszóhoz automatikusan kiválasztjuk a hozzá legközelebb álló másik szót egy előre létrehozott listából;
- „TOEFL teszt": speciális feleletválasztós teszt, ahol néhány alternatíva közül kell automatikusan kiválasztani a megadott szóhoz jelentésben legközelebb állót, TOEFL tesztfeladatot megoldva;
- szemantikai osztályozás (előre kijelölt kategóriákba, pl. gyümölcsök, fegyverek, stb.), szófaji és mondattani klaszterezés.

Ontológia alapú eljárások esetén szintén megjelent az irodalomban az automatikusan kiszámolt szóhasonlósági értékek összehasonlítása az emberi intuícióval, humán kísérletekre támaszkodva (Resnik 1995). Cikkem következő fejezete ilyen tanulmányok tapasztalatait ismerteti. A humán intuíció viszonyítási pontként való felhasználása *szintén felhasználható disztribúciós (korpuszalapú) vizsgálatokban* tesztelésre és értékelésre (v.ö. Dobó & Csirik 2012). A szakirodalomban a humán tesztelési módszer magyar adaptációja azonban eddig nem volt kidolgozva, e közlemény harmadik fejezete mutatja be az első ilyen vizsgálatot.

2 Humán szóhasonlósági mérések az angol szakirodalomban

Rubenstein és Goodenough (1965) kísérletében egyetemi hallgatók értékelték szókártyákon látható szópárok hasonlóságát. A kísérlet résztvevői először sorrendbe állították a kártyákat a rajtuk látható szópárok jelentésének hasonlósága szerint, majd osztályozták a hasonlóságukat 0-tól 4-ig, tetszés szerint törteket is használva. Munkájukban a szerzők a szinonímiára hivatkoznak, mint a megfigyelt hasonlóság okára, a kísérlet során azonban a szóhasználatuk más volt: a kísérlet résztvevőinek szóló instrukciók szerint a szavak „jelentésének hasonlóságát" kellett értékelni.

A kísérlet második részében a szerzők 100 résztvevőtől mondatokat elicitáltak ugyanezen szavak szerepeltetésével. A szavakat az így kapott mondatokban velük előforduló más szavak gyakoriságával jellemezték. A kísérletnek ez a része ma elavultnak számít, hiszen több százmillió szavas korpuszokat használunk hasonló célra. Érdemes figyelembe venni, hogy Rubenstein és Goodenough közleményének megírásakor még az egy millió szavas Brown korpusz sem volt elérhető, kísérleti eljárásuk mégis nagyon emlékeztet a mai számítógépes, nagy korpuszokon alapuló, disztribúciós szemantikai kutatások alapelveire. A szerzők méréseikkel és

statisztikai elemzés segítségével kimutatták, hogy a szópárok szubjektív szemantikai hasonlósága és a szavak megfigyelt környezete egymástól nem független.

Miller és Charles (1991) azt vizsgálták, hogy a szójelentés hasonlósága és a kontextus megkülönböztethetősége hogyan függ össze. A szójelentés hasonlóságát a fentiekkel megegyező módon, az 1965-ös kísérletből vett 30 szópár felhasználásával mérték, az eredeti módszerrel megismételve a kísérletet. A szópárokat úgy válogatták ki, hogy a szinonímia egyértelmű eseteiből tízet, a szinonímia nyilvánvaló hiányából újabb tízet, valamint további tíz, ebből a szempontból köztes esetnek számító szópárt választottak. Az 1965-ös kísérletben kapott átlagos hasonlósági pontszámokkal magas korrelációt mértek (Pearson r=0,97), annak ellenére is, hogy az egyik szópárt (valószínűleg egy egyszerű hiba folytán) megváltoztatták, (*cord, smile*) helyett (*chord, smile*) szerepelt a mérésükben.

A szavak hasonlóságának automatikus, objektív mérésében a WordNet lexikai adatbázis elkészítése áttörést jelentett (összefoglalóért ld. Meng, Huang & Gu 2013). A WordNetben a szavak szinonimahalmazokba vannak rendezve, melyeken belül a szemantikai hasonlóság maximális. A szinonimahalmazokat különböző szemantikai relációk kapcsolják össze. Ezekből elsősorban kettő, a hiponímia és a hiperonímia (alá- és fölérendeltség) használható fel legközvetlenebb módon a szemantikai hasonlóság mérésére: a fogalmak közötti szemantikai távolság mérhető a két fogalmat összekötő alá- ill. fölérendeltségi útvonalon bejárt csomópontok számával. Így pl. a *train* és a *public transport* szavakat tartalmazó szinonimahalmazok közti távolság egy lépés, a *boy* és a *male person* között szintén egy lépés a távolság; a *train* és *instrumentation* között három, a *boy* és a *being* között szintén három. Mivel egy *physical object* nevű csomópont mind a *being*, mind az *instrumentation* csomópontot dominálja, így megmérhető a *boy* és a *train* szavak közötti távolság is. Ennél természetesen összetettebb metrikák is léteznek, például olyanok, amelyek a fogalmak közti távolságot nem veszik konstans módon egy egységnyinek, továbbá a fogalmak információtartalmával is számolnak. Az információtartalom fordítottan arányos az előfordulás valószínűségével, melyet korpusz segítségével, az adott fogalom korpuszbeli előfordulásának gyakoriságával lehet mérni (gyakorlatilag viszont nem fogalmak, hanem szóalakok gyakoriságát mérik, mely persze egy jelentős leegyszerűsítése a problémának).

Resnik (1995) a szavak hasonlóságát a fogalmak WordNetbeli távolsága és a Brown korpusz segítségével számított információ-tartalom felhasználásával mérte. Módszere tesztelésére a Miller & Charles (1991) által alkalmazott (eredetileg Rubensteinéktől származó) humán kísérleti módszertant alkalmazta. Resnik tíz egyetemi hallgatóval ismételte meg a kísérletet, és megállapította, hogy a pontszámok szópáronkénti átlaga a saját kísérletében és a Miller-Charles kísérletben erősen korrelálnak (r=0,97). Ez hasonló érték ahhoz, amit Miller és Charles kapott 1991-ben. Resnik sem hagyta módosítás nélkül a szólistát: saját kísérletében a *woodland* szót tartalmazó szópárokat nem szerepeltette. Resnik automatikus módszerrel (WordNetből és korpuszból kapott információtartalomból) számolt szóhasonlósági értékei r=0,79 korreláció mellett együtt járnak az általa mért humán pontszámok átlagával (Resnik 1995).

A három kísérletben kapott eredményeket az 1. táblázat foglalja össze. Az 1965-ös és 1991-es adatsorok közti r=0,96 és az 1991-es és 1995-ös közötti r=0,97 korreláció alapján a módszer kísérletileg reprodukálhatónak és robosztusnak látszik. Ebből a szempontból érdekes lehet az a kérdés is, hogy egy-egy kísérleti alany eredményei hogyan korrelálnak az átlaggal. Resnik saját kísérletében az átlaggal való korreláció középértéke 0,88 volt, a szórás pedig 0,08, tehát ez a tesztelési módszer ebből a szempontból is megbízhatónak tűnik, miközben ez egyfajta emberi felső korlátnak is tekinthető ennek a feladatnak a megoldásában, melynél jobbat elvárni a gépi módszerektől sem érdemes.

		R & G (1965)	**M & C (1991)**	**Resnik (1995)**
car	automobile	3,92	3,92	3,9
gem	jewel	3,84	3,84	3,5
journey	voyage	3,84	3,58	3,5
boy	lad	3,82	3,76	3,5
coast	shore	3,60	3,70	3,5
asylum	madhouse	3,04	3,61	3,6
magician	wizard	3,21	3,50	3,5
midday	noon	3,94	3,42	3,6
furnace	stove	3,11	3,11	2,6
food	fruit	2,69	3,08	2,1
bird	cock	2,63	3,05	2,2
bird	crane	2,63	2,97	2,1
tool	implement	3,66	2,95	3,4
brother	monk	2,74	2,82	2,4
lad	brother	2,41	1,66	1,2
crane	implement	2,37	1,68	0,3
journey	car	1,55	1,16	0,7
monk	oracle	0,91	1,10	0,8
cemetery	woodland	1,18	0,95	-
food	rooster	1,09	0,89	1,1
coast	hill	1,26	0,87	0,7
forest	graveyard	1,00	0,84	0,6
shore	woodland	0,90	0,63	-
monk	slave	0,57	0,55	0,7
coast	forest	0,85	0,42	0,6
lad	wizard	0,99	0,42	0,7
cord* / chord**	smile	0,02 *	0,13 **	0,1 **
glass	magician	0,44	0,11	0,1
rooster	voyage	0,04	0,08	0,0
noon	string	0,04	0,08	0,0

1. táblázat: Angol nyelvű humán szóhasonlósági vizsgálatok eredményei (szópáronkénti pontátlagok)

3 Magyar nyelvű humán szóhasonlósági kísérlet

A rubensteini módszer szerinti humán szóhasonlósági vizsgálatot magyar nyelvre eddig nem végezték el. Dobó és Csirik (2012) az általuk kipróbált algoritmusok kiválasztásához és hangolásához használta ugyan tesztelési módszerként a Resnik kísérletben használt szókészlet egy magyar fordítását, ehhez azonban az angol (Rubenstein és Goodenough 1965) kísérletben mért hasonlósági adatokat párosították.

Az itt bemutatott kísérletben a Rubenstein és Goodenough-féle eredeti vizsgálatból Miller és Charles (1991) által kiválasztott szópárok saját fordítását készítettem el, majd 28 résztvevő[2]

[2] A mérésben ismerősök (10 fő), hallgatók (16 fő) és kollégák (2 fő) vettek részt. Az átlagéletkor 33 év volt, a nemek megoszlása: 14 férfi és 14 nő. Segítségüket ezúton is hálásan köszönöm!

bevonásával magyar nyelven is elvégeztem a kísérletet. Az előzményekben megfigyelhető *cord-chord* alternáció miatt 31 szópárral dolgoztam, így az adatokat mind a három angol nyelvű előzménykutatással össze tudtam vetni.

A kísérlet résztvevői számára az instrukciók a következők voltak:

1. Nézze meg az összes kártyát!
2. Jelentésük hasonlóságának sorrendjében rendezze őket csökkenő sorrendbe!
3. A szópárok utáni üres négyzetben osztályozza a hasonlóságuk mértékét 0-tól 4-ig. 0 a legkevésbé, 4 a teljesen hasonló. Adhat törtszámokat is osztályzatként (pl. 3,9). Több kártya is kaphatja ugyanazt az osztályzatot.

A kísérlet több résztvevője fontosnak találta elmondani, hogy mi alapján döntött (volt, aki általánosságban fogalmazva, és volt, aki konkrét szópárok kapcsán, mintegy magyarázatként). A szinonímia elsődleges döntési tényező volt. Elhangzott olyan megjegyzés – és ez a pontszámokon is megfigyelhető –, hogy ha a szinonímia segítségével nem dönthető el a sorrend, akkor asszociációkat keresnek a szavak között. Az egyik résztvevő megjegyezte, hogy ő ugyan elsősorban a szinonímiát, másodsorban a szavak közötti esetleges asszociációt kereste, ezek hiányában a szóalakok közötti hasonlóság is hatott rá. Az alá- és fölérendeltség hatását egyik résztvevő sem említette, de ez nem is jelent meg sok szópárban (kivétel: *madár-kakas* és *madár-daru*). Összehasonlításként: az ontológiát használó automatikus módszerek éppen alá-fölérendeltségi hierarchiákkal dolgoznak – és érnek el magas korrelációt ugyanezen tesztelési módszerrel kapott eredményekkel. Ennek egyik lehetséges magyarázata az is, hogy a szinonímia fogalmába intuitív módon bizonyos fokig az alá-fölérendeltségi helyzetet is beleértik a kísérlet résztvevői, legalábbis ennek a konkrét feladatnak a végrehajtása során. Tóth (2012) mérte a szinonímia mellett az antonímia és az alá-fölérendeltség megjelenését is a korpuszalapú eredményekben (saját tesztelési módszerrel), és ezeknek a kapcsolatoknak a hatása mind nagyon erősen kimutatható volt korpuszból automatikusan nyert hasonlósági értékek esetén is.

A hasonlóság megítélését befolyásolhatja a poliszémia vagy a homonímia, például a *testvér* szó esetén. Nem kaptak a résztvevők arra nézve utasítást, hogy ezeket az eseteket tudatosan keressék, és ha ilyet tapasztalnak egy szónál, akkor a kapott párjához közeli jelentését vizsgálják csak, vagy éppen ellenkezőleg: vegyék figyelembe az alternatív jelentéseket is. Egyfajta lexikális előfeszítést (*priming*) természetesen létrehozhat a szópár másik szava, meghatározva a kísérletben részt vevő emberben felmerülő olvasatot (pl. a *testvér* szó esetén a *szerzetes*), de éppen ennél a szópárnál látható (a maximális terjedelem eléréséből, és a magas szórásból), hogy ennek a tudat alatti folyamatnak a megbízható érvényesülésére ebben a kísérletben nem számíthatunk.

A kísérlet eredményeit a 2. táblázat mutatja be. Az egyes résztvevők válasza a kapott átlagos pontszámokkal magas korrelációt mutat (r=0,9), a szórás 0,04. Ez nagyon hasonló a Resnik által angol adatok alapján kapott értékekhez (r=0,88, szórás=0,08), amit ő egyben a humán teljesítmény elvárható felső korlátjaként aposztrofált.

A kapott *magyar* hasonlósági középértékek korrelációja az *angol* (eredeti) középértékekkel (Spearman-féle rangkorreláció) a következő:

- Rubenstein & Goodenough (1965) adataival r=0,959, p<0,01
- Miller & Charles (1991) adataival r=0,950, p<0,01
- Resnik (1995) adataival r=0,932, p<0,01

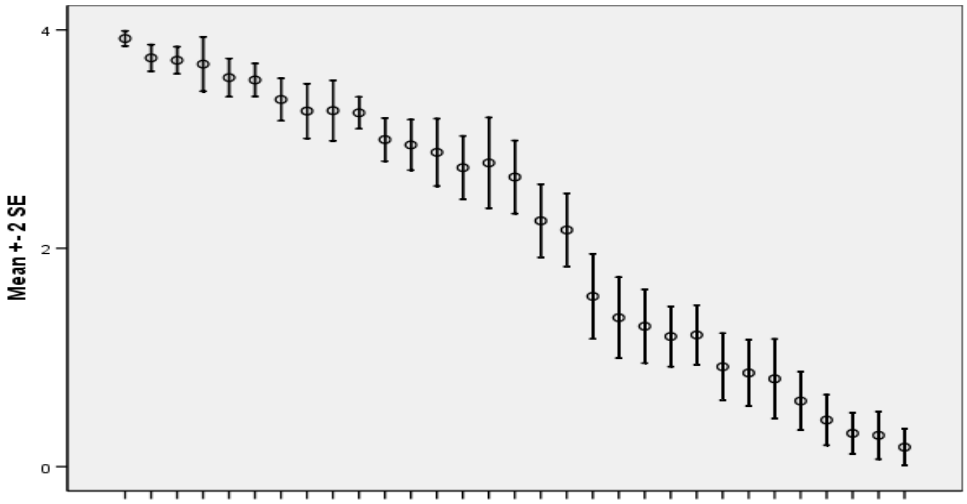
Mindhárom esetben erős, szignifikáns korrelációt tapasztalunk annak ellenére is, hogy egyrészt a magyarra fordítás során az eredetitől eltérő többértelműségeket hoztunk létre, másrészt ezek a szavak olyan területekre is elvezetnek minket, ahol a két nyelv szókincse eltérő gazdagságot mutat. Kiemelném azonban, hogy ez nem egy fordítási feladat, hiszen ezeket az adatokat éppen azért gyűjtjük, hogy különböző módszerekkel kapott *magyar* adatokat hasonlíthassunk össze egymással.

		hasonlóság átlaga	**terjedelem**	**szórás**
autó	személygépkocsi	3,92	0,8	0,18
drágakő	ékkő	3,74	1	0,32
dél	délidő	3,72	1	0,32
elmegyógyintézet	bolondokháza	3,69	3	0,65
fiú	legény	3,56	2	0,46
part	vízpart	3,54	1,6	0,39
szerszám	eszköz	3,36	2	0,50
kályha	tűzhely	3,26	3,4	0,65
bűvész	varázsló	3,26	3	0,72
utazás	út	3,24	1,4	0,38
étel	gyümölcs	3,00	1,9	0,51
madár	kakas	2,95	2,2	0,60
madár	daru	2,88	3,5	0,80
testvér	szerzetes	2,78	4	1,08
daru	eszköz	2,74	3,5	0,75
utazás	autó	2,65	3,3	0,87
étel	kakas	2,25	3,5	0,87
legény	testvér	2,17	3,3	0,87
part	domb	1,56	3	1,01
vízpart	erdő	1,37	3	0,96
part	erdő	1,29	3	0,88
erdő	sírkert	1,21	2,4	0,71
temető	erdő	1,19	2,3	0,72
szerzetes	jós	0,86	3	0,79
legény	varázsló	0,82	3	0,80
üveg	bűvész	0,80	3,6	0,95
szerzetes	rabszolga	0,60	3	0,69
húr	mosoly	0,43	2	0,60
kakas	utazás	0,31	1,7	0,49
fonal	mosoly	0,29	2,2	0,57
dél	kötél	0,18	1,8	0,44

2. táblázat: A magyar nyelven elvégzett humán szóhasonlósági kísérlet eredményei

A táblázatban, valamint az 1. ábrán is látható, hogy a szórás és a terjedelem a leghasonlóbbnak bizonyult szavak esetében volt a legkisebb. A hasonlóságra adott pontszámok átlagának középtartományában szóródtak az adataink a leginkább (jelezve a döntés bizonytalanságának növekedését). A legkevésbé hasonló szavak esetén újra lecsökkent a szórás, azonban már nem érte el a leghasonlóbb szavaknál tapasztalt értékeket.

Tóth Ágoston: Magyar nyelvű szóhasonlósági mérések humán és disztribúciós szemantikai kísérletben
Argumentum 9 (2013), 301-310
Debreceni Egyetemi Kiadó



1. ábra: A szórás változása – az x tengelyen a szópárok, hasonlóságuk átlagának csökkenő sorrendjében (a sorrend a 2. táblázatban látható)

4 A szubjektív szóhasonlósági adatok alkalmazása disztribúciós szemantikai kísérletben

Az előző szakaszban bemutatott humán szóhasonlósági mérés fő célja, hogy megfelelő eszközt adjon automatikus módszerek teszteléséhez, a működésüket meghatározó paraméterek beállításához.

Ennek kipróbálására a Tóth (2012) által bemutatott disztribúciós szemantikai eszköz kissé módosított változatával számítógépes szóhasonlósági mérést is végeztem. A Magyar Webkorpusz (Halácsy és mtsai. 2004) 100 millió szavas, lemmatizált (Trón és mtsai. 2005) alkorpuszával dolgoztam. A 2. táblázat első két oszlopában látható szavakat (a továbbiakban: *célszavak*) a környezetükben előforduló szavak (*környezetszavak*) gyakorisági adataiból képzett tulajdonságvektorokkal jellemeztem. Ez a vizsgálati szakasz együttes előfordulásra vonatkozó adatokat számszerűsít, szintagmatikus kapcsolatokat tár fel. A kísérletben a szavak figyelembe vett környezete csupán 1-1 szó volt mindkét oldalon. Minden egyes célszót egy annyi elemből álló vektor jellemez, ahány környezetszót használunk a vizsgálathoz: ebben a kísérletben a Magyar Webkorpusz 20000 leggyakoribb lemmájának előfordulását figyeltem a kiválasztott célszavak környezetében. Minden egyes előfordulás a vizsgált célszót jellemző vektornak az adott környezetszóhoz tartozó elemét növelte eggyel.

A környezetvektorok összeállítása után azok súlyozása következett. A vektorok elemeiből a cél- és környezetszavakra pozitív pontonkénti kölcsönös információt számoltam, ezzel mérve a két szó együttes előfordulásának valószínűségét azok külön történő előfordulásához képest.

A vizsgálat utolsó lépéseként a célszavakat jellemző vektorokat hasonlítjuk össze. A kiválasztott módszer a környezetvektorok hajlásszögének kiszámítása ($cos\ \alpha$) volt. Matematikailag ez egy geometriai művelet egy 20000 dimenziós vektortérben. Nyelvészeti szempontból az éppen összehasonlított két szó paradigmatikus kapcsolatait keressük.

A paramétereket (lemmatizálás használatát, vektorok súlyozását és összehasonlítását, a környezetszavak számát) kísérletezéssel állítottam be, a nagy számításigény miatt szuperszámítógépes környezetben dolgozva, saját fejlesztésű szoftver használatával. Néhány további

részlet leírása, valamint az eredmények egy alternatív módszer szerinti kiértékelése megtalálható Tóth (2012)-ben.

A korpuszból nyert szóhasonlósági adatokat a 3. fejezetben bemutatott humán kísérletben kapott eredményekkel hasonlítottam össze. A Spearman-féle rangkorreláció értéke r=0,591 ($p<0,01$), tehát a korreláció már 1%-os szinten is szignifikáns (a szubjektív humán és a korpuszban mért értékek együtt jártak), pozitív, közepesen erős.

Tóth (2012)-ben megoldandó feladatként és tesztelési eszközként egy feleletválasztós vizsgálatot használtam. Kiindulásként 15 szemantikailag motivált párt vettem: voltak köztük szinonimák (pl. *egész–teljes, fut–rohan, néz–figyel*), ellentétek (*fekete–fehér, régi–új, ki–be*) és hiponimák/hiperonimák (alá-/fölérendelt szavak, avagy specifikusabb/általánosabb szavak, pl. *alma–gyümölcs, labdarúgás–sport, szekrény–bútor, kutya–állat*), egyforma számban. A célszavak páronkénti megadása azt biztosította, hogy mindegyik szóhoz volt egy „legközelebbi szó", ami a rendszer által visszaadandó elvárt kimenet volt. A szavak kiválasztásánál a szófaji változatosság is szempont volt. A helyes kimenetet mind a 30 szóhoz összesen 100 potenciális jelölt közül kellett kiválasztania a rendszernek: a 100 alternatíva tartalmazta az eleve vizsgált 30 szót, valamint 70 olyan szót, amit a Magyar Webkorpusz első 1000 leggyakoribb szavából választott a program véletlenszerűen. A pontosság átlagos elvárható értéke véletlen választás esetén (*random baseline*) 1% volt. A fedést ennél a tesztelési módszernél 100%-on tartjuk: a feleletválasztás kikényszerített jellegű. A pontosság 20 millió szó feldolgozása után 79% volt (baseline: 1%); ezután már nem javult a pontosság, egészen 100 millió szóig vizsgálva.

A disztribúciós szemantikai kísérletsor későbbi folytatásában a paramétereket (pl. a figyelembe vett környezetet, a vektorok súlyozási és összehasonlítási módját) tervezem további kísérletezéssel úgy beállítani, hogy a kapott értékek a humán kísérlet eredményével minél jobban korreláljanak, figyelemmel kísérve ugyanakkor a Tóth (2012)-ben bevezetett tesztelési módszerrel kapott eredményeket is. A paraméterek változtatása során megfigyelhető, hogy a lemmatizálás bevezetése jelentős javulást hoz a magyar nyelv feldolgozásában (az angolban ez a hatás elmarad, v.ö. Bullinaria és Levy 2007). Ugyancsak nagy előrelépést hozhat a *mondattani elemzésen* alapuló súlyozás bevezetése; ezt egyelőre magyar nyelven, disztribúciós szemantikai kísérletben nem vizsgálták.

Dobó és Csirik (2012:219) saját mérései alapján *angol nyelvű* szövegek esetén – kizárólag korpuszadatokat felhasználva – r=0,77 az elért legmagasabb rangkorrelációs érték humán szóhasonlósági adatokkal összehasonlítva az automatikus algoritmusok által szolgáltatott adatokat. Resnik (1995) korábban bemutatott, WordNetet is használó rendszere r=0,81 korrelációt ért el Resnik saját mérésében kapott szubjektív szóhasonlósági adatokkal. Legutóbbi fejleményként létrehoztak a Resnik kísérletében kapott emberi „felső korlátot" megközelítő és elérő, nem (vagy nem kizárólag) korpuszadatokkal dolgozó rendszereket is (v.ö. Patwardhan és Pedersen 2006, Agierre és mtsai 2009).

5 Összegzés

A humán szóhasonlósági kísérletek (mind az angol szakirodalmi előzmények, mind a 3. szakaszban bemutatott magyar kísérlet) fő célja az, hogy más módszerekkel, objektív módon nyert szóhasonlósági értékek kiértékelését lehetővé tegyék, származzanak azok korpuszadatokból vagy más forrásokból. A szubjektív kísérlet résztvevőinek döntései nagyon hasonlóak voltak: a pontszámok szórása alacsony volt, az átlagként kapott pontszámokkal a rangkorreláció magas. A kísérlet résztvevői tapasztalataikat is megfogalmazhatták; ez értékes kvalitatív

visszajelzést is szolgáltatott a statisztikailag is kiértékelhető, kvantitatív eredményeken kívül. A magyar kísérletben kapott pontátlagok szignifikánsan és erősen korreláltak az angol kísérletekben kapott adatokkal is. Az eredmények azt mutatják, hogy ezek az adatok más eredmények kiértékelésére nagy biztonsággal felhasználhatók. Első ilyen felhasználásként elvégeztem a Tóth (2012) által bemutatott disztribúciós szemantikai algoritmus kiértékelését az új eszközzel is.

Irodalom

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Paşca, M. & Soroa, A. (2009): A study on similarity and relatedness using distributional and WordNet-based approaches. In: *10th Annual Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies. Association for Computational Linguistics*. Stroudsburg, 19–27.

Bullinaria, J.A. & Levy, J.P. (2007): Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods* 39, 510–526.

Dobó, A. & Csirik, J. (2012): Magyar és angol szavak szemantikai hasonlóságának automatikus kiszámítása. In: Tanács, A. & Vincze, V. (szerk.): *MSZNY 2013: IX. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 213–224.

Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I. & Trón, V. (2004): Creating open language resources for Hungarian. In: *Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004)*.

Meng, L., Huang, R. & Gu, J. (2013): A Review of Semantic Similarity Measures in WordNet. *International Journal of Hybrid Information Technology* 6(1), 1–12.

Miller, G.A. & Charles, W.G. (1991): Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6(1), 1–28.

Patwardhan, S. & Pedersen, T. (2006): Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. In: *11th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics,* Stroudsburg*,* 1–8.

Resnik, P. (1995): Using information content to evaluate semantic similarity. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence.* San Francisco, CA: Morgan Kaufmann Publishers Inc., 448–453.

Rubenstein, H. & Goodenough, J.B. (1965): Contextual correlates of synonymy. *CACM* 8(10), 627–633.

Tóth, Á. (2011): A multidisciplinary approach to lexical ambiguity. In: Mateoc, T. (szerk.): *Cultural Texts and Contexts in the English Speaking World*, 372–388.

Tóth, Á. (2012): Vektortér alapú szemantikai szóhasonlósági vizsgálatok. In: Tanács, A. & Vincze, V. (szerk.): *MSZNY 2013: IX. Magyar Számítógépes Nyelvészeti Konferencia.* Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 354–360.

Trón, V., Kornai, A., Gyepesi, Gy., Németh, L., Halácsy, P. & Varga, D. (2005): Hunmorph: open source word analysis. In: *Proceedings of the ACL Workshop on Software*. Ann Arbor, MI: Association for Computational Linguistics, 77–85.

Tóth Ágoston
Debreceni Egyetem
Angol-Amerikai Intézet
4010 Debrecen
Pf. 73.
toth.agoston@arts.unideb.hu