A beszélői szubjektivitás vizsgálata szentiment- és emóciókorpuszokon[i]

Drávucz Fanni[1], Szabó Martina Katalin[2, 3, 4]

[1] ELTE BTK Nyelvtudományi Doktori Iskola
[2] Szegedi Tudományegyetem, Szláv Intézet, Orosz Filológiai Tanszék
[3] Precognox Informatikai Kft.
[4] MTA TK „Lendület" RECENS Kutatócsoport
dravucz.fanni@gmail.com
szabo.martina@lit.u-szeged.hu

Kivonat: A dolgozatban a nyelvi szubjektivitás és az emóció- és szentimenttartalmak összefüggéseit vizsgáljuk számítógépes nyelvészeti eszközök és módszerek segítségével. A munka elméleti alapvetése az, hogy a bizonytalanságot jelölő elemek a beszélői szubjektivitás indikátorai lehetnek, ezért az elemzéshez a nyelvi bizonytalanságjelölők szólistáit alkalmazzuk. A munka során a bizonytalanságjelölő elemek szótáraival emóció- és szentimentkorpuszokat elemeztünk, és a különböző típusú szemantikai tartalmak gyakorisági összefüggéseit vizsgáljuk.Bár a jelenségekkel a nyelvtechnológiában a szentiment- és az emócióelemzés részben foglalkozik, nincs tudomásunk olyan dolgozatról, amely az emóciókat és a szentimenttartalmakat a nyelvi szubjektivitással összefüggésben, a nyelvtechnológia szempontjából górcső alá venné. A kutatásunk további haszna, hogy a nyelvi bizonytalanság detekciójára tesz kísérletet, amely feladat az automatikus értékeléselemzés kardinális problémája.

1 Bevezetés

A jelen dolgozatban a nyelvi szubjektivitás és az emóció- és szentimenttartalmak összefüggéseit vizsgáljuk számítógépes nyelvészeti eszközök és módszerek segítségével. A számítógépes nyelvészetben a szentimentek alatt a szerzői attitűdöt tükröző nyelvi elemeket (pl. „a főnököm remek ember"; Péter 1991), míg az emóciók alatt a szöveg szintjén tetten érhető érzelmeket értjük (pl. „nagyon örültem a főnökömnek"; Szabó et al. 2016a), melyeket a háttérben húzódó kognitív értékelő, illetve emotív funkciók különböztetnek meg.

[i] A jelen kutatást Az Emberi Erőforrások Minisztériuma Új Nemzeti Kiválóság Programja, az MTA Társadalomtudományi Kutatóközpont „Lendület" RECENS Kutatócsoportja támogatta, valamint a European Research Council (ERC) az Európai Unió „Horizon 2020" nevű, 648693-es számú kutatási és fejlesztési programja keretében. Az elemzések alapjául szolgáló egyes korpuszok, szótárak és adatok rendelkezésre bocsátásáért a szerzők köszönik Vincze Veronika segítségét.

Korábbi dolgozatunkban (Drávucz–Szabó–Vincze 2017) szentimentszótárakat alkalmazva emóciókorpuszt, és emóciószótárakat alkalmazva szentimentkorpuszt elemeztünk. Megállapítottuk, hogy a szentimentkorpuszunkban annotált értékelő tartalmaknak csupán a negyede volt jelezhető az emóciószótárainkkal. Úgy véljük, ez az eredmény egybevág azzal a megközelítési móddal, miszerint a szentimenteknekcsupán egy része tekinthető inkább szubjektív (pl. személyes, kontextusfüggő) ítéletnek, és az inkább objektív (pl. személytelen, tényszerű) ítéletek nem tartalmaznak emotív (ily módon emóciószótárral detektálható) elemeket. Mivel az emóciószótárral elemzett szentimentkorpusz (vö. Szabó–Vincze 2015, Szabó et al. 2016a) termékvélemény-szövegeket tartalmaz, valószínűsíthető, hogy a tesztelt termékek értékelői értékítéletek megfogalmazására, és nem saját érzelmi viszonyulásuk kifejezésére törekedtek bennük. E tapasztalat alapján úgy döntöttünk, megvizsgáljuk, vajon a szentimentek, az emóciók, valamint a beszélői szubjektivitás milyen összefüggései tárhatóak fel automatikus eszközökkel.

Az interakcióban megjelenő érzelmek, valamint a szubjektív tartalmak kifejezését a pszichológia tudománya részleteiben vizsgálja (vö. pl. Ekman–Friesen 1969, Ekman 2007), azonban a nyelvi elemek emotív illetve szubjektív tartalmának megismerésével a nyelvtudomány területén viszonylag kevés kutatás foglalkozik.

A jelenségekkel a nyelvtechnológiában részben a szentiment- és az emócióelemzés foglalkozik (vö. pl. Liu 2012, Mulcrone 2012, Szabó et al. 2016a, Szabó–Vincze–Morvay 2016), ugyanakkor nincs tudomásunk olyan dolgozatról, amely az emóciókat és a szentimenttartalmakat a nyelvi szubjektivitással összefüggésben, a nyelvtechnológia szempontjából venné górcső alá. A kutatásunk további haszna, hogy a nyelvi bizonytalanság detekciójára tesz kísérletet, amely feladat – amint arról később részletesen szólunk (l. fentebb) – az automatikus értékeléselemzés kardinális problémája.

2 Elméleti háttér

A jelen fejezetben tisztázzuk a kutatásunkban használt legfontosabb fogalmakat, illetve rámutatunk néhány, a terminológiát érintő fontosabb problémára.

A számítógépes értékeléselemzés, másképpen *szentimentelemzés* (*sentimentanalysis*) a számítógépes nyelvészet egy részfeladata, amely arra irányul, hogy az értékelő tartalmakat megtalálja a szövegekben, meghatározza ezeknek az értékeknek a típusát, azaz pozitív vagy negatív voltát, valamint, hogy megállapítsa az értékelés tárgyát, tehát azt, hogy az értékelés mire irányul (vö. Liu 2012). A számítógépes érzelemelemzés, másképpen *emócióelemzés* (*emotiondetection* vagy *emotionrecognition*) a szentimentelemzéstől eltérően tipikusan nem az adott dolog nyelvi értékelése, hanem a szövegekben megbúvó – adatközlő által megélt – emóciótartalom kinyerését célozza. Ily módon, bár a két tartalomelemzési feladat, sőt az azokon keresztül vizsgálható kognitív jelenségek is átfedést mutatnak, azok tárgya és célja nem azonos egymással (vö. Drávucz–Szabó–Vincze 2017).

A beszélői szubjektivitást a szentimentelemzés oldaláról megközelítve a következőket kell megemlítenünk: A nemzetközi gyakorlatban a szentimentelemzés feladatát gyakran a szubjektivitáselemzés fogalmával jelölik, e két terminust tulajdonképpen egymás ekvivalenseként használják (vö. pl. Wilson et al. 2005, Liu 2010). Ahogyan Liu (2010) érvel, a szöveges információk két alapvető kategóriára oszthatók: tényszerűekre és véleményekre. Ez utóbbiakat nevezi a szerző

szubjektív kifejezéseknek, amelyek a megnyilatkozó értékelését vagy érzését tartalmazzák valamely entitás, esemény, vagy azok valamely sajátsága vonatkozásában. Mindezek alapján a szerző nem tekinti a szentimentelemzés feladatához tartozónak a tényszerű közlések kivonatolását. Liuhoz (2010) hasonlóan, Wilson és munkatársainak szentimentelemző rendszere (Wilson et al. 2005) is a szubjektív mondatok azonosítását célozza, és dolgozatukban a feladatot szubjektivitáselemzésként említik.

Magunk a magyar nyelvű szövegek szentimentelemzésében nem törekszünk e két információtípus szétválasztására (vö. Szabó 2016a). Meglátásunk szerint a szentimentelemzés eredményének felhasználása szempontjából valójában nincs jelentősége annak, hogy feltárjuk, vajon egy adott értékelő kifejezés inkább tényszerű megállapítás, vagy inkább az adott megnyilatkozó szubjektív megítélése-e. Ebből a szempontból saját elemzési alapelvünk Péter (1991) osztályozási megoldásával áll összhangban, amelyet a nyelvi értékelés vonatkozásában tesz. A szerző megkülönbözteti az értékelés „racionális" és „emocionális" típusát egymástól (vö. Péter 1991: 46), és az előbbire *a főnököm remek ember*, az utóbbira *a habbeton rossz hővezető* mondatokat hozza példaként. Magunk tehát mindkét információtípus kinyerését célozzuk az automatikus értékeléselemzésünk során (vö. Szabó 2016).

A valódi jelentőséggel bíró sajátságot, és egyben problémát a szentimentelemzésben azt jelenti, hogy hogyan tudjuk kiszűrni azokat a nyelvi tartalmakat, amelyek esetében a megnyilatkozó a közölt információt illetően bizonytalan. A bizonytalanságot jelölő kifejezések automatikus azonosítása napjaink nyelvtechnológiai kutatásainak egyik fontos problémaköre (vö. Vincze 2014a: 99). A szövegek automatikus tartalmi elemzése során ugyanis csupán akkor tudunk megfelelő elemzési eredményt produkálni, ha az elemzőrendszerünk képes a tényszerűen közölt és a bizonytalan nyelvi tartalmakat elkülöníteni egymástól. A szentimentelemzés feladatában ez azt jelenti, hogy elsősorban azokat az értékelő tartalmakat kívánjuk kinyerni, amelyeket a megnyilatkozó tényként közöl. Tekintsük az alábbi példákat!

[1] A hangminőség jó.
[2] a. A hangminőség valószínűleg jó.
b. Jó a hangminőség?
c. Nem tudom, hogy a hangminőség jó-e.
d. A hangminőség jó lehet.
e. Minden bizonnyal jó a hangminőség.

Az [1] alatti, faktív olvasatú példával ellentétben a [2] alattiak nem faktív olvasatúak. Mindegyikük tartalmaz ugyanis valamilyen olyan nyelvi eszközt, amely lehetetlenné teszi az értékelést megfogalmazó szövegrész faktív olvasatát. Másképpen: megakadályozza, hogy az adott szentimentet a megfogalmazó által tényként közölt információként fogadjuk el (vö. Szabó–Vincze 2015: 221–222). Belátható tehát, hogy az automatikus értékeléselemzésben csupán az [1] alatti példa tartalmát volna szerencsés tényszerű értékítéletként feldolgozni, a [2] alatti példák tartalmát nem.

A probléma megoldásával az angol nyelv vonatkozásában több dolgozat is foglalkozik (vö. Farkas et al. 2010, Szarvas 2012). A magyar nyelvet illetően egyetlen kézzel annotált bizonytalansági korpuszról van tudomásunk, valamint az e korpusz feldolgozása alapján született első eredményekről a nyelvi bizonytalanságot jelölő elemek automatikus felismerését illetően (vö. Vincze 2014a) (az ebben a projektben alkalmazott osztályozási megoldásról részletesebben l. lentebb, [3]).

Amint arról a dolgozat bevezetőjében már szóltunk (l. fentebb), korábbi munkánkban (Drávucz–Szabó–Vincze 2017) szentimentszótárak segítségével emóciókorpuszt, és emóciószótárak segítségével szentimentkorpuszt elemeztünk. Megállapítottuk, hogy a kapott eredmények egybevágnak azzal a megállapítással, miszerint a szentimenteknek egy része objektívként klasszifikálható, amellyel összefüggésben azok nem tartalmaznak önreflexív elemeket. Ez a vizsgálati eredmény adta az alapötletét a jelen kutatásnak, amelyben a beszélői szubjektivitás, az emóciók, valamint a szentimentek összefüggéseit igyekszünk feltárni.

A szubjektivitás vizsgálatához a bizonytalanságot jelölő nyelvi elemeket választottuk. Úgy véltük ugyanis, hogy a bizonytalanságot jelölő elemek a beszélői szubjektivitás indikátorai lehetnek, ezért a bizonytalanság és a szentiment- és emóciótartalmak összefüggéseinek megismerésével valamelyest betekintést nyerhetünk a szubjektivitás és a szentiment- és emóciótartalmak összefüggéseibe is.

3 A munka módszere és eszközei

A nyelvi bizonytalanság automatikus felismerésére a természetesnyelv-feldolgozás (NLP) eszköztárából a szótárillesztés módszerét alkalmaztuk, amely például a gépi tanulás, a szintaktikai elemzésen alapuló mintaillesztés mellett egyszerűbb információkinyerési módszer. A kutatás során egy bizonytalanságjelölő elemeket tartalmazó szótárat (vö. Vincze 2014b) illesztettünk olyan magyar nyelvű korpuszokra, melyeket korábban kézzel annotáltunk az emóciók, illetve a szentimentek szintjén (vö. Szabó et al. 2016, Szabó–Vincze 2016).

Ahogyan arra Vincze (2014) felhívja a figyelmet, a nyelvi bizonytalanságot hagyományosan a mondat szemantikájához szokták kötni, azonban vannak olyanbizonytalanságot jelző nyelvi elemek is, amelyek a megnyilatkozás kontextusában –diskurzusbeli tényezőknek köszönhetően – válnak bizonytalanító értékűvé. Ezzel összefüggésben az alábbi példák nem azonos okból kifolyólag tekinthetőek tartalomelemzési szempontból bizonytalannak:

[3] a. Lehet, hogy esik az eső
b. Számos kutató szerint az ezüstkolloid belsőleg sokkal hatásosabb, mint a mesterségese antibiotikumok

Amíg a [3a] alatti példában azonosítani tudunk egy, a közölt tartalom faktivitását törlő nyelvi elemet, addig a [3b] alatti példa esetében ilyen elemet nem találunk; a bizonytalanságot az okozza, hogy semmilyen konkrét, ellenőrizhető információt nem kapunk a közölt tartalom forrására, hitelességére vonatkozóan. Azt látjuk tehát, hogy amíg egyes, bizonytalanságot kifejező mondatokban a bizonytalanság ténye valószínűleg automatikusan is könnyen azonosítható elemekhez kötődik, addig más esetekben az csupán a pragmatika szintjén érhető tetten, ami nyilvánvalóan az automatikus elemzés számára is komoly kihívást jelent (erről még a későbbiekben, az elemzés kapcsán is lesz szó, l. lentebb).

Szarvas és munkatársai (Szarvas et al. 2012), valamint Vincze (2013, 2014) dolgozataikban a szemantikai bizonytalanságnak négy, a diskurzusszintű bizonytalanságnak három fajtájával foglalkoznak részletesen. Magunk a jelen kutatásban erre az osztályozási rendszerre támaszkodunk. Az egyes típusokat az alábbiakban röviden, példák segítségével prezentáljuk.

I. Szemantikai bizonytalanság:

a. **Episztemikus** (**epistemic**): Egy propozíció episztemikusan bizonytalannak számít, ha a világtudásunk alapján nem tudjuk eldönteni az adott pillanatban, hogy igaz-e vagy hamis. Pl. *Lehet, hogy esik.*
b. **Feltételes mondatok** (**condition**): Az episztemikus bizonytalansággnál ismertetett sajátság jellemző az ún. hipotetikus bizonytalanságra is, amelyeket közé soroljuk a feltételes mondatokat (**condition**), pl. *Ha esik, itthon maradunk.*
c. **Vizsgálati bizonytalanság** (**investigation**): Nincs konkrét, ellenőrizhető információ a közölt tartalom forrására, hitelességére vonatkozóan, az például vizsgálattal lehet megállapítható Különösen tudományos cikkekben gyakori, hiszen a kutatási kérdést gyakran a vizsgálati bizonytalanság nyelvi eszközeivel fogalmazzák meg a szerzők. Pl.*A felvétel manipuláltságáról vizsgálatot folytattak.*
d. **Doxasztikus**(**doxastic**): A hiedelmekkel összefüggő, ebből adódó bizonytalanság. Pl. *Azt hiszi, hogy a Föld lapos.*

II. Diskurzusszintű bizonytalanság (az itt ismertetett típusokat – magyar nyelvű terminológia híján – angol nyelven nevezzük meg):

a. ***weasel***: az információtartalomhoz nincs egyértelműen forrás rendelve (nem derül ki, kihez köthető az adott információ), vagy pedig hiányzik a közlésből egy fontos és releváns információrészlet, amely szükséges lenne. Pl. *Egyesek szerint inkább megszállást kellene mondani.*
b. ***hedge***: homályossá teszik bizonyos mennyiségek vagy minőségek jelentését. Pl. *A belga lakosság kb. 10%-a él Brüsszelben.*
c. ***peacock***: bizonyítatlan (vagy bizonyíthatatlan) értékeléseket, minősítéseket vagy túlzásokat fejeznek ki. Pl.*Apafi négy évet keserves tatár fogságban töltött.*

A bizonytalanságjelölők szólistái, amelyeket a jelen kutatásban alkalmaztunk:

Típus	Elemszám
condition	6
doxastic	11
epistemic	13
investigation	7
hedge	37
weasel	26
peacock	22
Összesen	**122**

1. táblázat. A bizonytalanságjelölők elemszáma a szótárban, típusonként

Az emóciókorpusz szöveganyagát tévés és mozis témájú blogoldalakról származó, különböző terjedelmű és szerzőségű kritikákból, hírekből, valamint kommentekből állítottuk össze (vö. Szabó et al. 2016). A szentimentkorpuszunkat termékvélemény-szövegek alkotják (vö. Szabó–Vincze 2016). A korpuszokról néhány alapvető információt az alábbi táblázatban közlünk.

	Szentimentkorpusz	Emóciókorpusz
Méret	17 000 mondat 250 000 token	794 mondat 7205 token
Címkézés	bináris: pozitív/negatív	többcímkés: 7 emóció
Domain	termékvélemények	blog és komment
Forrás	divany.hu	tévés és filmes blogok
Példák	*A nem kávéfüggőknek bejött Omlós, finom, van benne valami nagyszerű.* *Ez a konkrét zacskó senkinek nem ízlett, túl száraz, túl pörkölt és túlságosan rossz.*	*Tesóm rámrúgta az ajtót már fél kilenckor nem örültem annyira…* *Közben persze egy csomó olyan cuccot találtam, amit fogalmam sincs, mikor táraztam be… :D* *Olyan csoda cuccokat találtam. :)*

2. táblázat. A felhasznált korpuszok alapvető adatai

A vizsgálatot a következőképpen végeztük el: először a bizonytalanságot jelölő elemek hét listáját illesztettük a szentiment- és az emóciókorpuszra, majd a kapott eredmények alapján megvizsgáltuk a különböző típusú bizonytalanságjelölők, valamint a különböző szentimentértékek és emóciók összefüggéseit.

4 Eredmények, következtetések

A vizsgálat legfontosabb feltételezése az volt, hogy a nyelvi bizonytalanság szoros összefüggést mutat a beszélői szubjektivitással, ily módon az emóciófragmentumokban magasabb frekvenciával fognak szerepelni, mint a szentimentfragmentumokban. Az elemzési eredményeket a jelen fejezetben ismertetjük.

4.1 Az emóciókorpusz elemzése a bizonytalanságjelölők szólistáival

Elsőként az emóciókorpuszban a valamely emóció címkéjével ellátott fragmentumok és bizonytalanságjelölő elemek illeszkedésének a gyakoriságát mutatjuk be.

Az emóciókorpuszunk összesen 794 fragmentumot és 7205 tokent (központozással)tartalmaznak. Ebből összesen 330 fragmentumban találtunk olyan kifejezést, amely bizonytalanság meglétére utal. Ez az összes token 4,6%-a, illetve az összes fragmentum 41%-a. A 3. táblázat közli a vizsgálat részletes eredményeit. A táblázatban a százalékos értékek azt mutatják, hogy a különböző típusú bizonytalanságjelölők hogyan oszlanak meg a különböző emóciófragmentumok között.

Típus	Össz.	Emóciófragmentum						
		düh	**feszült-ség**	**undor**	**félelem**	**öröm**	**bánat**	**meg-lepetés**
condi-tional	28	2 7,1%	4 14,3%	4 14,3%	0	6 21,4%	8 28,6%	4 14,3%
doxas-tic	48	4 8,3%	8 16,7%	0	4 8,3%	16 33,3%	8 16,7%	8 16,7%
epis-temic	16	2 12,5%	2 12,5%	0	2 12,5%	2 12,5%	6 37,5%	2 12,5%
investi-gation	2	0	0	0	0	2 100%	0	0
hedge	94	12 12,8%	22 23,4%	2 2,1%	2 2,12%	22 23,4%	28 29,8%	6 6,4%
weasel	110	14 12,7%	18 16,4%	2 1,8%	2 1,8%	38 34,5%	22 20,0%	14 12,7%
pea-cock	32	2 6,3%	4 12,5%	2 6,3%	2 6,3%	14 43,8%	6 18,8%	2 6,3%
Össz.	330	36 10,9%	58 17,6%	10 3,0%	12 3,6%	100 30,3%	78 23,6%	36 10,9%

3. táblázat. A nyelvi bizonytalanság különböző típusú elemeinek gyakorisága a 7 emóciónál (felül a szemantikai, alul a diskurzusszintű bizonytalanságjelölők)

A 3. táblázat adatai alapján a következő legfontosabb megállapításokat tehetjük: Az emóciófragmentumok közül a legkevesebb bizonytalanságjelölő az undor, a legtöbb pedig az öröm érzelemnél jelent meg. A legtöbb bizonytalanságjelölő elem az öröm mellett a bánat érzelemkategóriában jelent meg: majdnem minden harmadik bizonytalanságjelölő elemet az öröm, minden negyediket a bánat emóciótaggel annotált fragmentumban azonosítottunk. A bizonytalanságjelölők a tipikusan impulzív érzelmek (düh, undor, meglepődés) esetében jelentek meg ritkábban.

Ha a különböző érzelmek polaritását nézzük, tehát azt, hogy pozitív vagy negatív-e az adott emóció, akkor a bizonytalanságjelölők megoszlását tekintve a negatív polaritású emóciók vannak túlsúlyban.

A fejezet további részében áttekintjük a különböző típusú bizonytalanságjelölő elemek összefüggéseit a különböző emóciótípusokkal.

A bizonytalanságjelölők két fő típusa közül a diskurzusszintű bizonytalanságra utaló elemek fordultak elő nagyobb arányban.

A szemantikai bizonytalanságjelölők közül a doxasztikus, azaz a hiedelmekkel összefüggő típus fordult elő az emóciófragmentumokban a leggyakrabban, és az emóciótípusok közül a legnagyobb frekvenciával az öröm esetében szerepelt, pl. (A példák esetében az emotív elemet vastagon szedjük, a bizonytalanságjelölőt pedig aláhúzással jelöljük.)

[4] Kate még szerintem **üvöltözne** tovább

Módszerünk episztemikus típusú bizonytalanságjelölőket a bánat emóció fragmentumaiban azonosított kiemelkedően gyakran, pl.

[5] Laci Vilibának **panaszkodik** a félresikerült majdnem lánykérés miatt.

A vizsgálati bizonytalanságra utaló elem összesen két találata jól mutatja, hogy a jelenség a tudományos szövegek jellemzője, és az emóciókorpuszban feldolgozott szövegtípusra nem jellemző.

A diskurzusszintű bizonytalanságjelölő elemekről elmondható, hogy a weasel szerepelt a leggyakrabban az emóciófragmentumokban, és őt a hedge típusú bizonytalanságjelölő követte. A weasel típusú, azaz a forrás bizonytalanságát jelölő elemek legmagasabb számban az öröm emóciónál jelentkeztek, pl.

[6] egy kis **szerelmes** évődésbe kezdenek

Bizonytalanságjelölőket a félelemnél és az undornál ritkán azonosítottunk. A hedge típusú – mennyiségre és minőségre utaló bizonytalanságjelölő – elemek a bánat emóciónál voltak a leggyakoribbak, pl.

[7] **sajnos** összességében kevés olyan emlékezetes, kreatív ötletet integrált a forgatókönyvbe, amitől most egy közepesnél jobb műfajparódiáról beszélhetnénk

A peacock típusú, bizonyíthatatlan értékelést, minősítést vagy túlzást jelölő elemek az öröm emóciónál mutattak kiemelkedően gyakori előfordulást, pl.

[8] az egyik legnagyobb **élményből** maradsz ki.

4.2 A szentimentkorpusz elemzése a bizonytalanságjelölők szólistáival

A szentimentkorpuszösszesen251 202 tokent tartalmaz a 15 675 fragmentumban, melyek polaritása közel kiegyenlített: 8465 (54%) negatív,valamint 7210 (46%) pozitív polaritású. A szentimentkorpuszon módszerünk a bizonytalanságjelölők szótárával összesen 4553 esetben állapított meg bizonytalanságot, ez a tokenek 1,8%-a, illetve a fragmentumok 29%-a (l. részletesen a 4. táblázatot).

Típus	**Szentimentfragmentum**		**Össz.**
	negatív	**pozitív**	
conditional	163	125	288
doxastic	262	160	422
epistemic	154	77	231
investigation	0	0	0
hedge	1 521	1 207	2 728
weasel	212	96	308
peacock	254	322	576
Összesen	2 566 **30,3%**	1 987 **27,6%**	4 553 **29,0%**

4. táblázat. A nyelvi bizonytalanság különböző típusú elemeinek gyakorisága a szentimentfragmentumokban(felül a szemantikai, alul a diskurzusszintű bizonytalanságjelölők)

A különböző típusú bizonytalanságjelölők valamelyest eltérő megoszlást mutatnak a pozitív és a negatív szentimentfragmentumokban, és összességében jellemzően a negatív fragmentumokban gyakoribbak.

Amennyiben egyesével megvizsgáljuk a bizonytalanságjelölők különböző típusainak az előfordulását, a következő megállapításokat tehetjük: A bizonytalanságjelölők két fő típusa közül – akárcsak az emóciókorpusz esetében – itt is a diskurzusszintű bizonytalanságra utaló elemek fordultak elő nagyobb arányban. Ami az egyes típusokat illeti, meg kell említeni, hogy nem találtunk a vizsgálati bizonytalanságra utaló elemet. Nyilvánvaló, hogy ez a jelenség nem jellemző a termékvélemény-szövegekre, amelyekből a szentimentkorpuszt építettük. Mindhárom további, szemantikai típus a negatív fragmentumokban volt a gyakoribb.

A diskurzusszintű bizonytalanságjelölő elemek közül – meglepő módon az emóciókorpuszon tapasztaltaktól teljesen eltérően – a legmagasabb arányban a hedge típusú, mennyiségre és minőségre utaló bizonytalanságjelölő elem szerepelt.(A szentimenteket vastagon szedtük, a bizonytalanságjelölőt aláhúzással jelöltük.)

[10] a csomagolása elég **kecsegtető**

Ez a típus az emóciókorpusz fragmentumaiban a legkevésbé frekventált diskurzusszintű bizonytalanságjelölő volt. Mindemellett a weasel típusú elemek fordultak elő legritkábban a diskurzusszintűek közül, amelyek pedig az emóciókorpuszban a legnagyobb gyakorisággal szerepeltek ebből a típusból, például:

[11] Ez mégis valami **drágább** kategóriás lehet az olcsók között.

5 Összegzés

A dolgozatban a szubjektivitás, valamint az emóciók és a szentimentek összefüggéseit vizsgáltuk úgy, hogy a nyelvi bizonytalanságjelölő elemek szótáraival emóció- és szentimentkorpuszokat elemeztünk, és különböző típusú szemantikai tartalmak gyakorisági összefüggéseit kutattuk.

A vizsgálat eredményei alapján a következő legfontosabb megállapításokat tehetjük: Amíg az emóciókorpuszunkban az összes fragmentum 41%-ában (összesen 330 fragmentumban) találtunk olyan kifejezést, amely bizonytalanság meglétére utal, addig a szentimentkorpuszunkban valamivel kevesebben, a fragmentumoknak csupán a 36%-ában (összesen 5 600 fragmentumban). Más adatokkal, amíg az emóciókorpuszban az összes token 4,6%-át, addig a szentimentkorpuszban ugyanennek csupán az 1,8%-át azonosítottuk bizonytalanság-jelölőként. Mindez egybehangzónak látszik kezdeti hipotézisünkkel, miszerint a bizonytalanságjelölők szorosabb kapcsolatban állnak az érzelmekkel, mint a nyelvi értékeléssel, és ez gyakorisági megoszlásukban is megmutatkozik. Amennyiben helytálló az a feltételezésünk, hogy a bizonytalanságjelölők a nyelvi szubjektivitás egyfajta indikátorai, úgy a fentebbi vizsgálati eredmények összhangban állnak azzal a feltételezésünkkel is, miszerint a beszélői szubjektivitás is inkább az emóciókkal, mint a szentimentekkel mutat összefüggést.

Mind az emóció, mind a szentimentkorpusz elemzésekor azt láttuk, hogy a bizonytalanságjelölők a negatív polaritású emóciókban voltak túlsúlyban. Ez utóbbi

tapasztalat véleményünk szerint azt igazolja, hogy a negatív véleményünket árnyaltabban, kevésbé direkt módon fejezzük ki.

Az emóciókorpuszban a legtöbb bizonytalanságjelölő elem az öröm és a bánat érzelemkategóriában jelent meg. A bizonytalanságjelölők ritkábban jelentek meg a tipikusan impulzív érzelmek (düh, undor, meglepődés) esetében. A szemantikai bizonytalanságjelölők közül az episztemikus és hipotetikus típus a bánat, míg a doxasztikus az öröm emóció esetében volt a leggyakoribb. Ugyancsak az emóciókorpuszban a diskurzusszintű bizonytalanságjelölő elemek közül a weasel és a peacock típusúak az öröm emóciónál, a hedge típusúak pedig a bánat emóciónál a leggyakoribbak.

A vizsgálati bizonytalanságra utaló elemek mindkét korpuszunkban rendkívül ritkák voltak. Nyilvánvaló, hogy ez a nyelvi eszközkészlet nem jellemző a két korpuszban feldolgozott szövegtípusok egyikére sem.

A diskurzusszintű bizonytalanságjelölő elemek közül a szentimentkorpuszban – az emóciókorpuszon tapasztaltaktól eltérően – a legmagasabb arányban a hedge típusú (mennyiségre és minőségre utaló) bizonytalanságjelölő elem szerepelt, amely az emóciókorpuszban összességében is a második leggyakoribb típus volt. Mindemellett a weasel típusú elemek fordultak elő legritkábban a diskurzusszintűek közül, amelyek pedig az emóciókorpuszban a legnagyobb gyakorisággal szerepeltek ebből a típusból. A szentimentkorpuszban a diskurzusszintű bizonytalanságjelölő elemek közül a peacock a pozitív, a hedge és a weasel a negatív fragmentumokban volt gyakoribb.

A kutatás további tervezett lépéseként indokoltnak tartjuk jelen tapasztalatok alapján a bizonytalanságjelölő szótár, illetve az illesztési módszer továbbfejlesztését. Ehhez többek között más, egyéb szövegtípusokat is szerepeltető korpuszok elemzése szükséges. A bizonytalanság, érzelmek és szentimentek viszonyát a nyelvben a bizonytalanságkorpusz (Vincze 2014a, 2014b) emóció- és a szentimentszótárakkal való elemzésével tervezzük tovább vizsgálni.

Irodalom

Drávucz, F., Szabó M. K., Vincze V. 2017. Szentiment- és emóciószótárak eredményességének mérése emóció- és szentimentkorpuszokon. In: Vincze, V. (szerk.) *XIII. Magyar Számítógépes Nyelvészeti Konferencia. MSZNY 2017.* Szeged: JATEPress. 228–239.

Ekman, P., Friesen, W. V. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica.* 1: 49–98.

Ekman, P. 2007. *EmotionsRevealed*. New York: Times Books. Elérhető: https://zscalarts.files.wordpress.com/2014/01/emotions-revealed-by-paul-ekman1.pdf. Letöltve: 2017. október 6.

Farkas, R., Vincze V., Móra Gy., Csirik J., Szarvas Gy. 2010. The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010).* Association for Computational Linguistics: Uppsala. 1–12.

Liu, B. 2010. Sentiment analysis and subjectivity. In: Indurkhya, N., Damerau, F. J. (szerk.) *Handbook of Natural Language Processing*. Second Edition. Chapman and Hall/CRC. 627–666. Elérhető: https://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf. Letöltve: 2017. október 6.

Liu, B. 2012. *SentimentAnalysis and OpinionMining.* Kézirat. Elérhető: http://www.cs.uic.edu/~liub/FBS/SentimentAnalysis-and-OpinionMining.pdf. Letöltve: 2017. október 6.

Mulcrone, K. 2012. Detecting Emotion in Text. Elhangzott: UMM CSci Senior Seminar Conference, University of Minnesota, Morris, 2012. április 28.

Péter, M. 1991. *A nyelvi érzelemkifejezés eszközei és módjai.* Budapest: Tankönyvkiadó Vállalat.

Szabó, M. K., Vincze, V. 2015. Egy magyar nyelvű szentimentkorpusz létrehozásának tapasztalatai. In: Tanács, A., Varga V., Vincze V. (szerk.) *XI. Magyar Számítógépes Nyelvészeti Konferencia. MSZNY2015.* Szeged: JATEPress. 219–226.

Szabó M. K., Vincze V., Simkó K., Varga V., Hangya V. 2016a. A Hungarian Sentiment Corpus Manually Annotatedat Aspect Level. In: Calzolari, N., Choukri K., Declerck T., Goggi S., Grobelnik M., Maegaard B., Mariani J., Mazo H., Moreno A., Odijk J., Piperidis S. (szerk.) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016).* Párizs: European Language Resources Association (ELRA). 2873–2878.

Szabó M. K., Vincze V., Morvay G. 2016b. Magyar nyelvű szövegek emócióelemzésének elméleti nyelvészeti és nyelvtechnológiai problémái. In: Reményi, A. Á., Sárdi Cs., Tóth Zs. (szerk.) *Távlatok a mai magyar alkalmazott nyelvészetben*. Budapest: Tinta. 282–292.

Szarvas, Gy., Vincze V., Farkas R., Móra Gy., Gurevych I. 2012. Cross-genre and cross-domaindetection of semanticuncertainty. *Computational Linguistics*, 38: 335–367.

Vincze, V. 2013. Weasels, hedges and peacocks: Discourse-level uncertainty in wikipedia articles. In: *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. Nagoya: Asian Federation of Natural Language Processing. 383–391.

Vincze, V. 2014a. Bizonytalanságot jelölő kifejezések azonosítása magyar nyelvű szövegekben. In: Tanács, A., Varga V., Vincze V. (szerk.) *X. Magyar Számítógépes Nyelvészeti Konferencia*. *MSZNY2014.* Szeged: JATEPress. 99–108.

Vincze, V. 2014b. Uncertainty Detection in Hungarian Texts. In: *Proceedings of COLING 2014*. Dublin. 1844–1853. Elérhető: http://www.aclweb.org/anthology/C/C14/C14-1174.pdf. Letöltve: 2017. október 6.

Wilson, T., Hoffmann P., Somasundaran S., Kessler J., Wiebe J., Choi Y. et al. 2005. Opinion Finder: A system for subjectivity analysis. In: *Proceedings of HLT/EMNLP on Interactive Demonstrations*. Stroudsburg, PA: Association for Computational Linguistics. Elérhető: http://people.cs.pitt.edu/~swapna/papers/OpinionFinder-extendedabstract.pdf. Letöltve: 2017. október 6.