Nyelvtechnológiai módszerek a Budapesti Szociolingvisztikai Interjú lexikai és szintaktikai vizsgálatában

Váradi Tamás, Peredy Márta, Oravecz Csaba

MTA Nyelvtudományi Intézet
e-mail:{varadi,mperedy,oravecz}@nytud.hu

Kivonat A dolgozat célja a Budapesti Szociolingvisztikai Interjú társalgási moduljainak lexikai és szintaktikai elemzése nyelvtechnológiai módszerekkel. Az elemzés a gépi eljárással annotált szövegeket elsősorban statisztikai módszerekkel vizsgálja. A BUSZI társalgási nyelvhasználatát a Magyar Nemzeti Szövegtárból vett minta segítségével az írott nyelvhasználat jellemzőivel veti össze. Ahol erre mód nyílik, a BUSZI2 által vizsgált társadalmi csoportok közötti lexikális és mondatszerkesztésbeli nyelvhasználati különbségeket is vizsgálunk.

Kulcsszavak: beszélt nyelv, korpusz-összehasonlítás, korpuszhomogenitás, jellemzőszó-vizsgálat, mondatszerkesztés, szintaktikai elemzés

1. Bevezetés

A BUSZI 2 [7] öt foglalkozás szerinti társadalmi csoport nyelvhasználatát vizsgálja a szociolingvisztikai interjú Labov által kidolgozott módszerével. Ennek fontos eleme az irányított társalgás, melynek során a gondosan kiképzett terepmunkások kötelező, illetve tetszőlegesen választott témákat beszéltek meg az adatközlőkkel. A magnóra felvett anyag lejegyzése alapjában véve a helyesírási szabályokat követte, de a BUSZI vizsgálati kérdéseit tartalmazó szociolingvisztikai változók, illetve a beszéd prozódiai és paralingvisztikai kisérőjelenségei gondos megörökítését is. Az eredetileg házi norma szerint kidolgozott annotáció az anyag tartalmi felülvizsgálata után XML-szabványos alakra lett átalakítva.

A tanulmány két fő részre tagolódik. A 2. részben a lexikai vizsgálatok eredményeit mutatjuk be. A szokásos gyakorisági listák mellett, kísérletet teszünk a szövegváltozat egyedi jellemzőit tükröző lexikai mintázatok feltárására, valamint azok korszerű módszerrel történő vizualizációjára is. A 3. rész a szintaktikai elemzéseket tartalmazza, melyekhez az adatbázis reguláris lekérdező nyelvén definiált lokális grammatikákat használtunk fel. A szófajok és a felszíni szerkezeti minták statisztikai síkon megragadható jellegzetességeit az írott nyelvhasználattal, illetve a BUSZI2 adatközlő csoport egymás közötti összehasonlításával mutatjuk be. Rövid összefoglalás zárja a dolgozatot a 4. részben.

2. Lexikai vizsgálatok

2.1. Szókincs-gazdagsági vizsgálatok

A lexikai vizsgálatok legegyszerűbb változata a szövegek szókincsére irányul. Számos lehetséges mérőszám alkalmazható (pl. típus/token arány, hapax-gyakoriság, dislegomenon-gyakoriság), melyek több alkalmazásban is gyakran használatosak, például szerző-, illetve műfaj-azonosításban [5], de megbízhatóságuk éppen az egyszerűségük miatt alacsony. Az egyes szövegtípusok szókincsére vonatkozó néhány szembetűnő különbség azért kiolvasható belőlük. A mérőszámok közül néhány nagyon egyszerű statisztikát foglal össze a 2.1. táblázat. A kvóták kódjai az alábbi adatközlőcsoportokra vonatkoznak: KV1: tanárok; KV2: egyetemi hallgatók; KV3: bolti eladók; KV4: gyári munkások; KV5: szakmunkástanulók.

1. táblázat. Szóstatisztikai adatok a különböző szövegeken.

Jellemző	Korpusz						
	MNSz	Buszi	KV1	KV2	KV3	KV4	KV5
1. szóalak	224128	173331	36846	29278	40994	37116	29097
2. szótípus	52876	26449	8971	6776	8639	8601	6560
3. típus/token	0.236	0.1526	0.2435	0.2314	0.2107	0.2317	0.2255
4. normált szóalak	25000						
5. szótípus	10140		6704	6048	5866	6283	5935
6. típus/token	.4056		.2682	.2419	.2346	.2513	.2374
7. főnév	6813		4109	3535	3299	3808	3070
8. ige	3904		4231	3869	4544	4362	4396
9. Fn/Ige	1.7451		.97116	.91367	.7260	.8729	.6983
10. Hapax	4402		2416	2082	1912	2189	2021
11. Dislegomenon	1014		564	538	522	577	548

A 3. sor magasabb típus/token aránya abszolút mértékben gazdagabb szókincset tükröz (többféle szó fordul elő adott nagyságú szövegben), viszont a korpusz növelésével a típusok száma nem nő arányosan, ezért a normált korpuszméretből (25 ezer szó) számított érték (6. sor) mutatja pontosan az írott és beszéd változat közötti eltérést ezen mutató tekintetében. Jól látható, hogy az MNSz szövegeit szignifikánsan magasabb érték jellemzi. Szembetűnő az eltérés a főnév/ige használatban is, itt a beszélt nyelvi szövegre mutatható ki egyértelműen az igék használatának magasabb aránya a főnevekhez képest. Az egyszer, illetve kétszer használatos szótövek (10., 11. sor) előfordulási gyakoriságának különbsége is egyértelműen jelzi a írott változat nagyobb lexikális gazdagságát.

Fontos megjegyezni, hogy ugyan a kvóták között is jelentkezik különbség a mérőszámokban, megbízható eredményekhez azonban részletesebb vizsgálatokra, illetve nagyobb mennyiségű szövegre lenne szükség.

2.2. Jellemzőszó-vizsgálatok eredményei

Számos lehetséges módszer közül (l. pl. [3]) az alábbiakban egy olyan eljárás eredményeit mutatjuk be, amely az egyes korpuszok gyakorisági profiljainak összehasonlításával határozza meg az adott szövegre jellemző lexikai elemeket. Ebben az összehasonlításban azok a nyelvi elemek szerepelnek a rangsor elején, amelyek a két összehasonlított korpuszban jellegzetesek, mindig a másikhoz viszonyítva, vagyis az eljárás egy közös listát generál, melyet utána kvalitatív vizsgálatnak lehet alávetni.

A vizsgálatban először a két korpusz nyers gyakorisági listáit állítjuk elő, majd minden, a listában szereplő szóra log-likelihood statisztikát számolunk [4]. Az így kapott eredmények szerint rendezzük újra a gyakorisági listát, így a lista elején megkapjuk az egyik vagy másik korpuszra jellemző szavak halmazát.

Az alábbi táblázatokban szereplő listákban az első oszlop a számított súlyérték, a második a szó(tő), harmadik az egyik (C1), illetve másik (C2) korpuszbeli gyakorisági érték.

1. MNSz vs. teljes Buszi

6143.08381829616	hát	C1: 107	C2: 4232
3341.37775731983	igen	C1: 118	C2: 2512
3273.87396188346	én	C1: 307	C2: 2991
2688.3186152523	nem	C1: 2873	C2: 6672
2277.40930049444	van	C1: 2274	C2: 5438
1962.18484387733	a	C1: 21240	C2: 9687
1435.46798174574	szóval	C1: 21	C2: 973

2. KV1 vs. KV5

326.999757436891	hát	C1: 376	C2: 1044
91.7885470809332	akko	C1: 0.5	C2: 70
84.3856052479682	meg	C1: 146	C2: 347
68.5496070494782	szóval	C1: 46	C2: 162

3. KV5 vs. KV1

88.3354097204393	gyerek	C1: 108	C2: 12
53.6296019468509	ugye	C1: 59	C2: 5
49.7205434485945	a	C1: 1550	C2: 1182
47.4355156652772	gimnázium	C1: 40	C2: 1
42.7362191531778	tanít	C1: 50	C2: 5
42.0983293946033	tanár	C1: 36	C2: 1

Az eljárás eredményei elnagyolva, de nagyon szemléletesen ábrázolhatók „szófelhők” formájában, melyet az 1. ábra illusztrál.

1. ábra. Teljes Buszi vs. MNSz.

2.3. Korpuszok homogenitásának és hasonlóságának vizsgálata

Korpuszok hasonlóságának vizsgálata, illetve mérése leginkább nyelvtechnológiai alkalmazások kontextusában merül fel és arra vonatkozóan adhat hasznos információt, hogy meghatározott típusú szövegekre fejlesztett alkalmazások milyen ráfordítással alakíthatók át újabb szövegek kezelésére. Ha a szövegek hasonlóak, a ráfordítás feltehetően kisebb. Általánosságban ennél többet nem is nagyon lehet állítani, egyrészt a lehetséges mértékek is nagyon sokfélék, másrészt megbízhatóságuk is jelentősen függ az adott szövegektől, illetve alkalmazásoktól.

A hasonlóság kérdése szervesen összefügg egy adott korpusz homogenitásának kérdésével, hiszen nyilván nem egyértelmű, mit fejezne ki egy ilyen mérték abban az esetben, ha egy homogén korpuszt hasonlítunk össze egy általános, nagy változatosságot mutató heterogén korpusszal (a kérdés részletes tárgyalását l. [3]). Ideális esetben ugyanazt a mértéket lehet használni korpuszon belüli és korpuszok közötti „távolságok” mérésére, melyek így közvetlenül összehasonlíthatók. A jelen vizsgálatokban alkalmazott perplexitásmérték használható ilyen módon, de természetesen, mint minden más egydimenziós mérték, csak durva közelítésként értelmezhető egy alapvetően többdimenziós összehasonlítási feladatban.

A perplexitás annak mérésére szolgáló érték, hogy mennyire pontosan tudunk modellezni egy ismeretlen valószínűségi eloszlást a rendelkezésre álló tanító adatok alapján. Minél magasabb értéket kapunk a perplexitásra, annál bizonytalanabbak vagyunk az ismeretlen eloszlás tekintetében, annál kevésbé tudjuk megjósolni az ismeretlen eloszlás által generált adatokat. Szövegek vizsgálatára az alábbi módon tudjuk felhasználni: a szöveg egy meghatározott részét (pl. 90%) használjuk fel tanító adatként, melyből például egy trigram (szóhármas) alapú statisztikai nyelvi modellt építünk (vagyis szóhármasok előfordulási gyakorisága alapján számoljuk ki a mondatok valószínűségét). Ezek után lemérjük, hogy a készült modell alapján milyen pontosan tudjuk megjósolni a tesztadatként (a maradék 10%) szereplő szövegrész mondatait. Minél nagyobb perplexitásértéket kapunk, annál távolabb áll a tanító és tesztszöveg egymástól, minél kisebbet,

annál homogénebb és annál jobban hasonlít a két szövegrész egymásra. Informálisan, a perplexitásra kapott számérték annak a szóhalmaznak a nagyságát határozza meg, amelyből (trigram nyelvmodell esetén) a megelőző két szó ismeretében a következő szót választhatjuk. Minél kisebb ez a halmaz, modellünk annál megszorítottabb [2].

2. táblázat. Perplexitásértékek a különböző szövegeken.

Tanító korpusz	Tesztkorpusz						
	MNSz	Buszi	KV1	KV2	KV3	KV4	KV5
1. MNSz	733.618	–	–	–	–	–	–
2. Buszi	–	121.52	–	–	–	–	–
3. KV1	–	–	123.835	123.462	113.633	118.273	107.597
4. KV2	–	–	122.666	115.782	–	–	–
5. KV3	–	–	124.402	–	101.542	–	–
6. KV4	–	–	130.106	–	–	108.828	–
7. KV5	–	–	127.512	117.237	110.695	116.796	89.401

Az itt végzett vizsgálatok sztenderd tízszeres keresztvalidációval készültek, a CMU-Cambridge Statistical Language Modeling Toolkit [1] segítségével, a morfológiai variábilitásból eredő eleve magas értékeket kiküszöbölendő a szokásos gyakorlatnak megfelelően szótövesített szövegekkel. A kapott eredmények a 2.3. táblázatban láthatók. Az egyes sorokban szereplő szövegekből készült a nyelvmodell, az oszlopok jelzik a tesztadatot. Abban a cellában, ahol mindkét, a sorban és oszlopban szereplő szöveg azonos, ott az adott korpusz homogenitására vonatkozó érték szerepel, a további cellákban pedig a különböző korpuszok hasonlóságát jellemző érték jelenik meg. Mivel a vizsgálat illusztratív, nem törekszik kimerítő jellemzésre, inkább a szembetűnő jellegzetességekhez kíván kvantitatív mérőszámot rendelni, ezért nem minden cellában szerepel (az egyébként minden esetben számítható) mutató. Néhány összehasonlítás a szövegek jellegéből következően nem hordoz lényeges információt, így azokat eleve nem érdemes elvégezni. Mivel az itt szereplő MNSz-minta jól láthatóan igen heterogén, nagy variábilitású szövegeket tartalmaz, a Buszi-szövegekkel való összehasonlítás nem eredményezne újabb információt azon túl, hogy az írott szöveg a beszélthez képest sokkal változatosabb, ez pedig a homogenitásadatokból is egyértelműen látszik már. A Buszi-szövegek vizsgálatában pedig informatívabb az egyes kvóták anyagát egymással összehasonlítani, mint a teljes Buszi-anyagot a kvóták anyagával; ez utóbbi esetben sem kapunk az előbbi vizsgálathoz képest új információt.

Az egyes korpuszrészek, kvóták homogenitására vonatkozó értékből kiolvasható, hogy az adott kvótához tartozó beszélőknek mennyire változatos a nyelv-

használata. A kvóták egymással történő összehasonlításából kapott értékek arra adnak választ, hogy a kvóták szövegei mennyire állnak közel egymáshoz, illetve az egyik szöveg milyen mértékig „foglalja magában" a másikat. A KV1 és KV2 korpusz például ebben az összehasonlításban viszonylag távol esik egymástól, míg ha a KV5 korpuszhoz hasonlítjuk például a KV1 korpuszt, akkor jelentős távolságot kapunk, fordított irányban pedig alacsonyat, vagyis a KV1 korpuszból épített modell „magában foglalja" a KV5 korpuszt is.

3. Szintaktikai elemzések

3.1. Mondathossz

A szintaktikai vizsgálatok alapegysége a mondat, így minden szintaktikai elemzés a mondathatárok megállapításával kell, hogy kezdődjön. Az írott nyelvi korpuszban ez nem jelent problémát, a beszélt nyelvi korpuszt tanulmányozva azonban talán a korpusz elemzésének legbizonytalanabb pontja éppen ez. A beszélők ugyanis (szemben az írott szövegek létrehozóival) nem jelzik egyértelműen, hogy hol van szerintük a mondataik vége. A BUSZI-korpusz tagolásánál a szöveget annotáló személyek anyanyelvi intuíciójuk alapján állapították meg a mondathatárokat.

A két korpusz közti első szembetűnő különbséget a 3. táblázat mutatja. Az írott nyelvi anyag átlagos mondathossza (17,1 szó) kétszerese a BUSZI-adatközlők élő beszédbeli mondatainak (8,5 szó). A BUSZI-terepmunkások megszólalásainak célja elsősorban az adatközlők beszédének terelgetése volt, így nem meglepő, hogy az ő megszólalásaik még rövidebb mondatokra tagolódnak. (A teljes BUSZI-beli átlagos mondathossz 6,5 szó.)

3. táblázat. Átlagos mondathossz.

	BUSZI		MNSZ
	terepmunkások (tm)	adatközlők (ak)	
átlagos mondathossz	4,6	8,5	17,1

Ragozott igealakok – tagmondatok. A mondatszerkezet szempontjából a legfontosabb eltérés a ragozott igealakok számában figyelhető meg. A BUSZI-ban másfélszer annyi ragozott ige van (15%), mint az MNSZ-ben (10%), l. alább 4. táblázat. Ez az adat utal arra az alább alaposan vizsgált tényre, hogy az írott nyelv több információt sűrít a főnévi csoportokba jelzős szerkezetek segítségével, míg a beszélt nyelv több alárendelt mondatot, és így több ragozott igét használ. Figyelembe véve, hogy tagmondatonként egy ragozott igével számolhatunk, megállapítható a tagmondatok átlagos hossza. A BUSZI-ban 6,7, az MNSZ-ben

10 szó adódik. Ezeket az értékeket összevetve a feljebb említett átlagos mondathosszal (BUSZI: 6,5; MNSZ: 17,1) azt kapjuk, hogy a BUSZI mondatai jellemzően egy tagmondatból állnak, hiszen az átlagos mondat- és tagmondathossz gyakorlatilag azonos, míg az MNSZ mondatai 1,7-szer hosszabbak, mint a tagmondatai, tehát a tipikus mondat két tagmondatból áll.

A bővítmények száma. Az NP-k számát a tagmondatok számához (azaz a ragozott igékhez) viszonyítva, azt látjuk, hogy míg a BUSZI-ban kettőnél kevesebb NP jut egy tagmondatra, addig az MNSZ-ben 3,5, vagyis az írott nyelv mondatai több bővítményt tartalmaznak. (L. alább 3.1. pont és a 4. táblázat.)

3.2. Szófajstatisztika

Már a legdurvább statisztikai elemzés, a különböző szófajú szavak számának összevetése is sokat elárul a beszélt nyelvi és az írott nyelvi korpusz mondatszerkezeti különbségeiről. A 4. táblázat a különböző szófajú szavak megoszlását mutatja a két korpuszban. Láthatjuk, hogy a legtöbb esetben az adatközlők és a terepmunkások szófajarányai közel azonosak még a diskurzusban betöltött eltérő szerepek ellenére is, míg az írott nyelvi szófajmegoszlás jelentősen eltér. Megjegyezzük, hogy a dolgozatban alább közölt statisztikai eltérések, ha külön nem jelezzük, akkor 5%-os szignifikanciaszint mellett mindig szignifikánsak.

4. táblázat. Szófajok.

Szófaj	BUSZI				MNSZ	
	tm		ak			
	%	$\sum$	%	$\sum$	%	$\sum$
N-ek száma	12,5	11904	14,0	24345	29,0	87479
Pro-k száma	13,7	13013	12,9	22388	5,5	16667
számnév	2,3	2200	3,7	6435	3,6	10861
egy-ek száma	1,0	917	1,4	2457	0,6	1930
Det-ek száma	6,4	6094	7,0	12058	12,3	37135
A-k száma	5,6	5315	5,6	9649	10,7	32149
Adv-ok száma	20,5	19533	20,0	34722	7,6	22879
finit V-k száma	15,2	14477	15,3	26570	9,9	29943
mn-i igenevek	0,5	512	0,5	946	3,3	9840
fn-i igenevek	1,7	1616	1,7	3010	1,0	3096
hat-i igenevek	0,2	146	0,2	331	0,3	896
kötőszók	10,9	10393	12,2	21086	7,5	22651
névutó	0,7	634	0,9	1567	1,6	4778
indulatszó	1,5	1461	0,4	644	0,0	116
egyéb	7,3	6919	4,1	7120	6,9	20881

3.3. A főnévi csoport

Az alábbiakban a főnévi csoportok szerkezetével foglalkozunk részletesebben, ugyanis a közölni kívánt tartalom átadásának két véglete közül az egyik az, amikor minden egyes információdarabnak egy-egy tagmondat felel meg, míg a másik véglet a tömörített szöveg, amelyben az információ minél nagyobb részét egy mondatba kívánja foglalni a beszélő (vagy a szöveg írója), és ezért a tartalom jelentős része a mondaton belüli főnévi csoportokban jelzői szerkezetekbe sűrítve jelenik meg.

A beszélt és az írott nyelvi korpusz főnévi csoportjainak összehasonlításakor fő hipotézisünk tehát az, hogy az írott nyelvben sokkal inkább megfigyelhető az információ főnévi csoportokba tömörítése, mint a beszélt nyelvben.

A főnévi csoport feje. A főnévi csoport feje főnév vagy névmás lehet és megfordítva, minden főnévre, illetve névmásra épül egy teljes főnévi csoport. Az 5. táblázatban a főnévi csoportok számát a főnevek plusz névmások számával azonosítottam, ami annyiban pontatlan, hogy a jelzőkkel bővített főnévi csoportból olykor el van hagyva a főnévi fej, illetve a mutató névmás nem mindig alkot önálló főnévi csoportot (pl. *ezt a kutyát*). Ezekről az esetekről alább még lesz szó. A főnévi csoportok jellemzően a mondat ragozott igéjének bővítményeiként jelennek meg a mondatban, de melléknévi csoportok (pl.: *büszke a fiára*), más főnévi csoportok (pl.: *a fiú a távcsővel*, *a fiúnak a távcsöve*) és igenevek (pl.: *a kertben játszó gyerek, uszodában úszni*) bővítményei is lehetnek.

Összességében több főnévi csoport van az MNSZ-ben, mint a BUSZI-ban. A főnevek és névmások összesített aránya a teljes szószámhoz képest rendre 35%, illetve 26%. Ez az adat máris mutatja, hogy az írott nyelvi korpuszban nagyobb szerepe van a főnévi csoportoknak, mint a beszélt nyelvben, összhangban azzal a 3.1. pontban említett adattal, hogy a ragozott igék relatív száma viszont a beszélt nyelvben magasabb.

Fontos további jellemzője a beszélt nyelvi korpusznak, hogy a főnévi csoportok között sokkal nagyobb arányban vannak a névmások, mint az írott nyelvben. Míg az írott nyelvben a félreértés elkerülése végett érdemes egy teljes leírással egyértelműsíteni, hogy mire utalunk, addig a beszélt nyelv sokkal inkább támaszkodhat az egyértelműsítés nem nyelvi eszközeire is (pl. mutatás), illetve esetleges félreértés esetén lehetőség lenne visszakérdezni, így a figyelem középpontjában álló (széliens) individuumokra elegendő csupán névmással utalni. A főnévi, illetve névmási fejek aránya a BUSZI-ban közelítőleg 50-50%, míg az MNSZ-ben 84-16% a főnevek javára.

Az adatokhoz három pontosító megjegyzést kell fűznünk. Egyrészt meg kell jegyeznünk, hogy a jelzővel bővített NP főnévi feje olykor elmaradhat (pl. *a sárga tulipánból* helyett *a sárgából*), a nem alanyesetű melléknevek csak ilyen esetekben jelennek meg, ezért az esetragos melléknevek és főnevek számának összevetéséből látható, hogy milyen gyakran maradhat el a főnévi fej a főnévi csoportokból. A BUSZI-ban ez az arány 7,2%-nek adódik a terepmunkások és 6,4%-nek az adatközlők esetében, míg csupán 3,5% az MNSZ-ben. Az ellipszisek valódi száma

5. táblázat. Főnévi csoportok.

	BUSZI		MNSZ
A főnévi csoport feje	tm	ak	
főnevek aránya (%)	47,8	52,1	84,0
névmások aránya (%)	52,2	47,9	16,0
A főnévi csoportok száma			
a szószámhoz képest (%)	26,2	26,6	34,5
a finit igék számához képest (%)	1,7	1,8	3,5

azonban ennél alacsonyabb, ugyanis bizonyos főnévként és melléknévként is érthető szavak melléknévként vannak megjelölve a korpuszban, és ezért például az *a törpéket* főnévi csoportban a *törpe* esetragos melléknévként számolódik. Az ebből fakadó hiba vélhetőleg egyformán érinti a BUSZI és az MNSZ korpuszt, így ha a kapott értékek nem is pontosak, arányuk jól mutatja, hogy az MNSZ NP-i teljesebbek, nemcsak hogy ritkábban fejezhetők ki névmással, de a főnévi fej is kevésbé hagyható el belőlük.

Másrészt, mint említettük az NP-k, bár leggyakrabban a mondat ragozott igéjének bővítményei, de nem feltétlenül azok, és ezek az esetek torzítják az egy ragozott igére eső NP-k számára kapott értéket. Harmadrészt a mutató névmások (*ez*, *az*) összes előfordulásainak a BUSZI-ban mintegy 20%-a, az MNSZ-ben 27%-a nem önálló NP-ként, hanem egy határozott főnévi csoporttal együtt fordul elő, ezeket tehát le kell vonnunk az önálló főnévi csoportként elszámolt névmások közül. Ez a kis korrekció azonban a névmások és főnevek arányára kapott értékeket lényegében nem módosítja.

Jelzős szerkezetek. Feltevésünk szerint az írott nyelvi korpuszban több és összetettebb jelzős szerkezeteket találunk, mint a beszélt nyelvben. Ezt vizsgáljuk alább a névelőt is tartalmazó NP-ken a melléknévi, majd a melléknévi igeneves jelzők esetén.

Halmozott melléknévi jelzők

A BUSZI-ban a névelős főnévi csoportoknak kb. 58%-a bővítetlen, az MNSZ-ben hasonló, de ennél valamivel alacsonyabb, 54% az arány. Az egy melléknévi jelzőt tartalmazók közel kétszer annyian vannak az MNSZ-ben, mint a BUSZI-ban, a két melléknévvel bővítettek már 2,5-szer, a hárommal bővítettek négyszer annyian. Négy melléknévi jelzőt tartalmazó NP a BUSZI-ban már nem található.

Melléknévi igenevek

A melléknévi igenevek használata sokkal gyakoribb az MNSZ-ben, mint a BUSZI-ban, az adatokkal azonban óvatosnak kell lennünk, mert a melléknévi igenevek közül sok valójában már melléknévként lexikalizálódott (pl.: *elvált*), elkülönítésükre azonban az annotáció nem ad lehetőséget.

A jelző + főnév szerkezetek között a melléknévi igenévi jelző a BUSZI-ban kb. 11%-ban, míg az MNSZ-ben kétszer olyan gyakran, 22%-ban fordul elő.

6. táblázat. A bővítetlen és a melléknevekkel bővített névelős főnévi kifejezések százalékos aránya a névelők összes számához képest.

	BUSZI		MNSZ
Halmozott mn.-i jelzők	tm	ak	
névelő+főnév	60,0	57,3	54,2
ne+mn+fn	8,60	8,77	17,07
ne+2mn+fn	0,90	0,86	2,36
ne+3mn+fn	0,06	0,06	0,23
ne+4mn+fn	0,00	0,00	0,01

7. táblázat. A melléknévi igenevek százalékos aránya a szavak számához viszonyítva.

	BUSZI				MNSZ	
M. igenevek	tm		ak			
	%	$\sum$	%	$\sum$	%	$\sum$
folyamatos	0,3	293	0,3	600	1,9	5806
befejezett	0,2	214	0,2	340	1,3	3922
beálló	0	5	0	6	0	112

8. táblázat. A melléknévi és melléknévi igenévi jelzők aránya.

	BUSZI		MNSZ
Igenévi/melléknévi jelzők	tm	ak	
melléknévi igenév+fn	11,2	10,1	21,8
melléknév+fn	88,8	89,9	78,2

A melléknévi igenevek használata jó módja az információ NP-n belüli tömörítésének, mivel az ige nemcsak magában, hanem bővítményeivel együtt is megjelenhet így jelzőként. A 9. táblázat adatai alátámasztják azt a feltételezést, hogy az írott nyelvi szöveg sokkal inkább él ezzel a lehetőséggel, ugyanis mintegy négyszer olyan gyakran van bővítménye az igenévi jelzőnek az MNSZ-ben, mint a BUSZI-ban.

9. táblázat. A bővítményes melléknévi igenévi jelzők százalékos aránya melléknévi igenévi jelzők között.

	BUSZI				MNSZ	
Bővített mn.-i ign. jelzők	tm		ak			
	%	$\sum$	%	$\sum$	%	$\sum$
bővítmény+m. igenév+fn	11,5	25	10,5	43	41,3	2271

Birtokos szerkezet. A kétféle birtokos szerkezet, az alanyesetű, illetve a -nAk ragos birtokost tartalmazó, megoszlása eltér a két korpuszban. A birtokot közvetlenül megelőző, nem névmási birtokost tartalmazó szerkezeteket (pl. *a kutyának a szőre* vs. *a kutya szőre*) vizsgálva kitűnik, hogy a -nAk ragos (datívuszos) birtokos aránya a beszélt nyelvben lényegesen nagyobb. A BUSZI-adatközlők között több, mint hússzor gyakoribb, mint az MNSZ-ben (10. táblázat). Ez szintén arra utal, hogy az írott nyelv sokkal inkább a tömörségre törekszik: ha az adott szerkezet egyértelmű, akkor fölösleges a kitett raggal redundánsan megjelölni a birtokost. A kötetlen beszéd kevésbé "spórol". Továbbá egy közel kétszeres szorzókülönbség a terepmunkások és az adatközlők adatai között is mutatkozik.

10. táblázat. A birtokot közvetlenül megelőző birtokosok között a datívusz aránya.

	BUSZI				MNSZ	
Birtokos szerkezetek	tm		ak			
	%	$\sum$	%	$\sum$	%	$\sum$
alanyesetű birtokos	89,6	146,81	5,211	99,1	5101	
részesesetű birtokos	10,4	17,18	5,48	0,9	44	

Többszörös birtokos szerkezetek a BUSZI-ban nem fordulnak elő, míg az MNSZ-ben igen, a birtokos szerkezetek 6%-ában.

3.4. Vonatkozói mellékmondatok

A BUSZI korpusz azt bizonyítja, hogy az *amely* és a *mely* vonatkozó névmás a beszélt nyelvből mára szinte teljesen eltűnt. Az adatközlők közül a *mely*-t senki, az *amely*-t csak a tanárok és az egyetemisták használták, így az adatközlők által használt összes vonatkozó névmásnak csak 0,7%-a volt *amely*, míg az MNSZ referenciakorpuszban a *mely* és az *amely* együttesen 41%-ot tesz ki. Az *amelyik* viszont az írott nyelvből hiányzik (0,4%), míg a BUSZI-ban több, mint 3%-ot képvisel. A beszélt nyelvben ugyanis az *amelyik* nem csak kiválasztó értelemben szerepel, hanem az *ami* és az *amely* helyett is, l. (1). A vonatkozó névmások BUSZI-beli használatáról részletesen ír [6].

(1) *s az Árpád Gimnáziumnak akkor még volt egy öööö nagyon jól működő cserkész csapata, amelyik különböző rendezvényeket öööö gyártott, rendezett*

A vonatkozó névmások összes száma megadja a vonatkozói mellékmondatok számát. A vonatkozói mellékmondatok összes NP-hez képesti aránya az MNSZ-ben alacsonyabb, 2,9%, míg a BUSZI-ban a terepmunkások esetében 4,1%, az adatközlőknél 4,3%, vagyis a beszélt nyelv feltevésünknek megfelelően, valóban gyakrabban fogalmazza a mondanivalót külön tagmondatba.

11. táblázat. A vonatkozó névmások előfordulásai.

Vonatkozó névmások	BUSZI tm %	BUSZI tm ∑	BUSZI ak %	BUSZI ak ∑	MNSZ %	MNSZ ∑
aki/ami-k száma	92,4	970,96	1,1978	58,8	1799	
amely-ek száma	4,2	44,0	6,13	33,1	1011	
mely-ek száma	0,3	3,0	0,0	7,7	236	
amelyik-ek száma	3,1	33,3	3,67	0,4	12	

A vonatkozói mellékmondatok közül csak a mondatkezdő pozícióban állókat vizsgálva szintén érdekes különbségek adódtak a két korpusz között. A vonatkozói mellékmondatok topikalizációval kerülnek a mondat élére. Ha a vonatkozói mellékmondatnak névmási feje van a mondatban, akkor ez a névmás mindig a mutató névmás (*az*). Az MNSZ-ben azonban a vonatkozói mellékmondat névmási feje az esetek felében el van hagyva, pl. (2-a), míg a BUSZI-ban szinte mindig megjelenik, pl. (2-b).

(2) a. ***Aki*** *erre jár és körül akar nézni ,* ~~***azt***~~ *szívesen fogadjuk* (MNSZ)
b. ***Aki*** *hisz Istenbe,* ***az*** *hisz pap nékül is* (BUSZI)

A 12. táblázat első két sora mutatja ezt az eredményt. A terepmunkások szövegében összesen 8 olyan mondat fordult elő, amelybe a mondat eleji vonatkozói mellékmondat után beilleszthető a mellékmondat *az* névmási feje, és ebből csupán egyszer maradt el az *az*, míg az adatközlők esetében 19 esetből egyszer sem. Ezzel szemben az MNSZ-ben 43 esetből 21-ben el volt hagyva az *az*. A névmás hiányát tekinthetnénk az elhagyott hangsúlytalan személyes névmási fej esetének, ám ekkor az itteni eredmények ellentmondanának a *hogy*-os tagmondatok fejével kapcsolatban tapasztalt tendenciának, mely szerint az írott nyelv sokkal inkább a hangsúlyos *az* névmást használja, míg a beszélt nyelvben gyakrabban előfordul fejként a hangsúlytalan személyes névmás is, és ez utóbbi lenne az, ami (alany- és tárgyesetben) elhagyható. Az adatok helyes értelmezése az, hogy a vonatkozói mellékmondatot már önmagában, a névmási fej nélkül is referáló bővítményként tudjuk értelmezni, és ekkor a névmási fejre nincs szükség. A mondatkezdő vonatkozói mellékmondat után mégis gyakran és legfőképpen a beszélt nyelvben megjelenő *az* inkább topikismétlő névmásnak tekinthető.

A mondatkezdő vonatkozói mellékmondatoknak a következő csoportja a visszautaló típus, lásd a 12. táblázat 3. sorát. A BUSZI-ban ugyanis a mondat eleji vonatkozói mellékmondat gyakran nem az adott főmondat valamely frázisának bővítménye, hanem az előző mondat valamely szereplőjére utal vissza, például (3). Ezekben az esetekben a vonatkozó névmás szintén referenciális kifejezésként viselkedik, tulajdonképpen személyes névmási funkcióban jelenik meg. A terepmunkások beszédében a mondatkezdő vonatkozói mellékmondatok 27%-a, az adatközlők beszédében ezek 42%-a utalt előző mondatbeli szereplőre, míg az MNSZ-ben csupán 9%.

12. táblázat. A mondatkezdő vonatkozói mellékmondatok típusai.

Vmm kezdetű mondat	BUSZI tm		BUSZI ak		MNSZ	
	%	∑	%	∑	%	∑
vmm + *az*	21	2,7	38,0	19	31,9	22
vmm + elhagyott *az*	3,0	1	0,0	0	30,4	21
visszautaló vmm	27,3	9	42,0	21	8,7	6
kettőspontos értelmezés	3	1	0	0	13	9
egyéb	45,5	15	20	10	15,9	11

(3) *Előtte Zuglóban laktunk a nagymamáméknál. Aki most ott lakik szintén egyedül*

Végül az MNSZ-ben több példát is találunk (13%) a mondatkezdő vonatkozó mellékmondat kettőspontos értelmezésére, pl. (4), míg a BUSZI-ban összesen egy ilyen mondat szerepelt. A kettőspontos értelmezés az *Ami . . . , az az, hogy . . .* mondat rövidített változatának tekinthető, ezt a tömörítést jellemzően az írott nyelv alkalmazza.

(4) *Ami még ennél is fontosabb: a televíziók nem a háború valóságos emberi vonatkozásaira voltak kíváncsiak*

4. Összefoglalás és további feladatok

A tanulmányban az írott és beszélt nyelvhasználat néhány jellemző különbségét illusztráltuk lexikai és szintaktikai elemzés alapján. Viszonylag egyszerű eszközökkel kaptunk nem triviális eredményeket, melyek alapul szolgálhatnak további nyelvi elemzéseknek.

A beszélt nyelvi korpusz méretének és a mondatelemzés mélységének a növelésével részletesebb vizsgálatok is elvégezhetők, mint például a BUSZI kvóták közötti különbségek nyelvstatisztikai elemzése vagy a szórendre vonatkozó elemzések.

Hivatkozások

1. Clarkson, P. R. és Rosenfeld, R. Statistical language modeling using the CMU-Cambridge toolkit. In: *EUROSPEECH-97*, 1. kötet, 1997, 2707–2710.
2. Jelinek, F., Mercer, R. L., Bahl, L. R. és Baker, J. K. Perplexity – a measure of the difficulty of speech recognition tasks. *Journal of the Acoustical Society of America*, November, 1977, 62:S63. Supplement 1.
3. Kilgarriff, Adam. Comparing Corpora. *International Journal of Corpus Linguistics*, 2001, 6(1):1–37.

4. Rayson, Paul és Garside, Roger. Comparing Corpora using Frequency Profiling. In: *Proceedings of the Workshop on Comparing Corpora*. Association for Computational Linguistics, 2000., 1–6.
5. Stamatatos, Efstathios, Fakotakis, Nikos és Kokkinakis, George. Automatic Text Categorization in Terms of Genre and Author. *Computational Linguistics*, 2000, 26 (2):471–495.
6. Szeredi, Dániel. Vonatkozó névmások használata beszélt nyelvi korpusz alapján. Szakdolgozat, ELTE, 2008.
7. Váradi, Tamás. A Budapesti Szociolingvisztikai Interjú. In: Kiefer, Ferenc és Siptár, Péter szerk. *A magyar nyelv kézikönyve*. Akadémiai Kiadó, Budapest, 2003, 339–359.