Fordítástudomány XII. (2010) 2. szám 67–95.

Kollokációk korpuszalapú vizsgálata

Reményi Andrea Ágnes
E-mail: remenyi@colbud.hu

Kivonat: A számítógépes korpuszok lendületet adtak a gyakran előforduló, több szóból álló szemantikai egységek gyűjtésének, elemzésének. E tanulmány elsősorban kvantitatív és nyelvtani kritériumok alapján definiálja ezeket, és csak másodsorban, az alosztályok képzésére használ szemantikai kritériumokat. Egy lemma két magyar nagykorpuszban (Magyar Történeti Korpusz, Magyar Nemzeti Szövegtár) előforduló kollokációinak bemutatása során sorra veszi azokat a kvantitatív elemzési lehetőségeket, melyek segítségével a gyakori szócsoportok közül kiemelhetők a lexikográfiában jelentősek.

Kulcsszavak: kollokáció-kutatás, osztályozási kritériumok, Magyar Történeti Korpusz, Magyar Nemzeti Szövegtár

1. A több szóból álló szemantikai egységek definíciója, osztályozása

A hagyományos szótárak a gyakran együtt előforduló szavak közül elsősorban a lexikográfiai szempontból legjellegzetesebbekkel, az állandósult szókapcsolatokkal foglalkoznak. A modern lexikográfiában és a számítógépes nyelvészetben azonban lehetőség van az összes, gyakran egymás közelében álló szópár (*bigram*), szóhármas stb., más szóval *szótöbbes* (*N-gram*) kigyűjtésére, de ehhez a definíciós hangsúlynak át kell kerülnie a tisztán szemantikai kritériumokról a könnyebben formalizálható mennyiségi és grammatikai kritériumokra. A magyar lexikográfia a szótöbbesek közül a legtöbb figyelmet az idiómáknak, vagyis az állandósult szókapcsolatoknak szenteli. Így O. Nagy Gábor (1966/1982)[1] és Ittzés Nóra (2002)[2] is a **korlátozott kompozícionalitás**ban találta meg ezek legfontosabb jellemzőjét, vagyis abban, hogy a kifejezés jelentése nem következtethető ki az őt alkotó elemek eredeti jelentéséből. Ittzés (2002) emellett utal arra is, hogy az elemek nem cserélhetők másra, és a köztük lévő szintaktikai viszony sem alakítható át. Ezek a definíciók tehát a kötött forma mellett a jelentés alapján klasszifikálnak, akárcsak a tudomásom szerint első nagyobb terjedelmű, kifejezetten a „visszatérő szókombinációként" (VII), ill. „fix kombinációként" (IX) definiált kollokációkkal foglalkozó Benson-Benson-Ilson szótár (1986)[3], mely az idiómákon kívül a másik oldalon kizárja a szabad kombinációkat is.

Geart Van Der Meer (1998: 314) a fentiekhez hasonlóan szemantikai alapon sorolja három csoportba a gyakran együtt előforduló szavakat. Kettős kritériumrendszerét a pre-konstrukcióra és a szó szerinti jelentéstől való eltérésre alapozza: „*a szabad kombinációk* nem pre-konstruáltak, és szemantikailag szó

szerint értendők (vagyis a komponensek megtartják konvencionális, szó szerinti jelentésüket). A *kollokációk* pre-konstruáltak, és szemantikailag szó szerint értendők (vagyis a komponensek megtartják konvencionális, szó szerinti jelentésüket). Az *idiómák* pre-konstruáltak, és szemantikailag nem szó szerint értendők (vagyis a szavak – vagy legalábbis az egyik szó – nem tartják meg szó szerinti, konvencionális jelentésüket, vagy legalábbis nem elemezhetők akként)." A nehezen formalizálható pre-konstruáltságot hangsúlyozza Hausmann (1985, idézi Heid: 2002) osztályozása is ("késztermék/félkésztermék" [*Fertigprodukte*/*Halbfertigprodukte*]).

Az előbbi megközelítésekkel ellentétben a számítógépben tárolt és sokszempontú, programozható kereséssel elemezhető korpuszok a formalizálható kritériumok előtérbe kerülésének kedveznek. Egy szövegkorpusz számítógépes feldolgozása esetén ugyanis lehetőség van arra, hogy egy program segítségével az összes közvetlenül egymás mellett álló, vagy egy, kettő, három, … közbeékelődő szó által elválasztott szótöbbest kigyűjtsük (az adott szóalakok, vagy morfológiai szempontból elő-elemzett korpusz esetén – s ez látszik célravezetőbbnek – a szótövek, vagyis a lemmák szintjén), és ezek közül gyakoriságuk vagy alkotórészeik statisztikai mutatókkal mérhető kapcsolati szorossága alapján kiválasszuk azokat az **adatjelölték**et (*candidate data*), melyeket azután a szemantikai kritériumok mentén csoportosíthatunk.

A gyakran együtt előforduló szavak, vagyis a szótöbbesek tehát három kritérium szerint definiálhatók:

- kvantitatív kritérium: a szótöbbesek nem egyszeriek, hanem visszatérőek. (A kollokáció azért fordul elő gyakran, mert lexikai elem, és nem fordítva [ld. Van der Meer 1998: 316]). Egy adott korpuszban megállapíthatunk egy küszöbértéket, amely feletti relatív gyakoriságot (vagy más statisztikai mutató szerinti értéket) elérő szótöbbeseket kiemeljük. (Olyan mérőszámot érdemes választani, amely a gyakoriságot a lemmán belül és a korpusz egészéhez képest is értelmezi.) Ezzel kizárjuk a „véletlenül" egymás mellé kerülő elemeket, vagyis a nem gyakori szabad kombinációkat;
- nyelvtani kritérium: a szótöbbesek szófaji szempontból jellegzetes, többnyire két vagy három tagú, egy szintaktikai szerkezetbe tartozó lexikai egységek, főnévből és/vagy igéből és/vagy melléknévből és/vagy határozószóból, valamint esetleg zárt szófajú szavakból (névmás, névelő stb.) állnak;
- A szemantikai kritériumok a szótöbbesek alosztályait különböztetik meg egymástól, így más-más csoportba kerülnek a *számot ad (valamiről)*, *példát ad (valamiről*, pl. *bátorságról)* (versus *példát mond*), *pénzt ad* kifejezések.

  A. Az idiomaticitás (vagyis a kompozícionalitás hiánya, amikor az egyik vagy mindegyik tag eredeti jelentéséből nem komponálható meg a közös jelentés), valamint két másik jellemző tulajdonság (a tagok behelyettesíthetőségének és a szintaktikai változtatásnak a tilalma) segítségével az állandósult szókapcsolatok, közmondások stb. csoportját választhatjuk le (*számot ad* [*valamiről*]). Ezek a szótöbbesek a hagyományos és a modern egynyelvű szótárakban is szerepelnek;

B. Az idiomatikus jelentéssel nem, de valamilyen hozzáadott jelentésmomentummal azért rendelkező szótöbbesek közül érdemes különválasztani azokat, melyekben a potenciális szinonimák közül többnyire csak egy alcsoport társulhat a szótöbbes másik/többi tagjához. Ezek egy része nem triviális, vagyis nem megjósolható. Ezeket a nem is idiomatikus, de nem is triviális együttes előfordulásokat nevezem **kollokáció**knak (*példát ad/példát mond*), a tagjait pedig **kollokáltak**nak. (A kollokációnak azt a tagját, amelynek alapján az adott elemzés folyik, gyakran **csomópont**nak (*node, base*) nevezik – ilyenkor a másik tag neve *kollokált*.) Így a *sovány-vékony* szinonimapár egymással felcserélhető, vagyis szabadon kombinálható az *arc, testalkat* főnevekkel, ugyanakkor komplementáris a *vigasz*, *eredmény, sugár, réteg, szelet* főnevek esetében: a *vigasz, eredmény* főnévhez többnyire a *sovány* melléknév, a *sugár, réteg, szelet* főnevekhez a *vékony* melléknév járul.

sovány arc/vékony arc
sovány teremtés/vékony teremtés
sovány vigasz/?vékony vigasz
sovány eredmény/*vékony eredmény
*sovány sugár/vékony sugár
*sovány réteg/vékony réteg
*sovány szelet/vékony szelet

A kollokációkat a hagyományos szótárírás nem tüntette fel az egy- és többnyelvű szótárak szócikkeiben, az újabb szótárak azonban egyre több figyelmet szentelnek ennek a lexikai dimenziónak. Például az idegennyelv-tanulók számára e szemantikailag transzparens formák a megértésben ugyan nem jelentenek problémát, de a szabad kombinálhatóság esetenkénti tilalma a produkcióban gondot okozhat a számukra;

C. A szótöbbesek fenti két típusára jellemző valamilyen hozzáadott jelentésmomentum. Bizonyos nyelvészeti felhasználások számára azonban szükség lehet a triviálisnak tekinthető, kompozícionális, szabad kombinációk ismeretére is (*pénzt/munkát/ösztöndíjat ad*), ha ezek gyakoriak. Például egy gépi fordítóprogram számára nincsenek triviális, magától értetődő szópárok, a gyakran előforduló szabad kombinációknak ezekben a felhasználásokban *van* információértékük.

Fontosnak tartom megjegyezni, hogy a szemantikai kritérium nehezen formalizálható volta miatt egyes szótöbbesek osztályozása igen nehéz.

A gyakori szótöbbesek összegyűjtésekor érdemes különválasztani a számítógépesíthető és a nem gépesíthető munkafázisokat. Ha a számítógépen tárolt korpusz morfológiailag elemzett, a kvantifikálás, vagyis a szótöbbesek lemma-szintű kigyűjtése és az őket jellemző statisztikai értékek kiszámítása gépileg megoldható.

Ha a korpusz szintaktikai szempontból nem elemzett, és mégsem csak a közvetlenül egymás mellett álló szavakra akarjuk szűkíteni a keresést, akkor a közbeékelődő szavak számát mechanikusan kell megadnunk; az angol nyelvre

végzett kutatások tapasztalatai (pl. Sinclair 1991, 1997) szerint az optimum maximum négy közbeékelődő szó. Ez az ún. **kollokációs ablak** (*span*), amelyen belül az összes lehetséges szótöbbesre kiszámolja a program a megfelelő mutatókat. Ez a mechanikus távolságon alapuló keresés sokkal magasabb hibaarányt fog eredményezni, hiszen például a mondat-, tagmondat- vagy főnévi vagy igei csoport-határ két oldalán talált szótöbbeseket is találatként fogja megadni. Minél mélyebb a szintaktikai elemzés, annál kisebb lesz a hibaarány a találatok között. Szintaktikai elő-elemzés hiányában a kollokációs ablak használatából következő hibaarányt úgy csökkenthetjük, ha az adatjelölteket eresztjük át egy szófaji szűrőn (Justeson és Katz 1995, idézi Manning és Schütze 1999: 143-4).

S végül a találatok közti, szemantikai kritériumok szerinti válogatás egyelőre gépileg nem oldható meg – ez a nyelvész feladata marad.

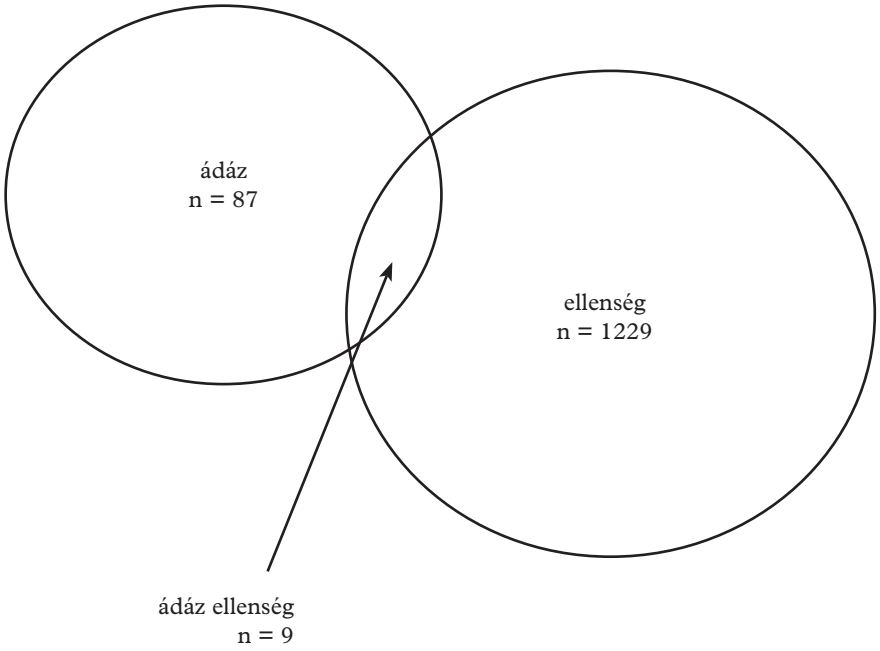
Az alábbiakban egy lemma két magyar nagykorpuszból a kollokációs ablak módszerével géppel gyűjtött, de kézzel elemzett szótöbbeseit mutatom be. Az elemzésen túl igyekszem bemutatni a további gépesíthetőség lehetőségeit és módszereit, ezzel próbálva egy első választ adni a Pajzs Júlia (2000: 217) által feltett kérdésre, hogy van-e automatizálható módszer a szókapcsolatok felismerésére, ill. az angol korpuszalapú lexikográfiában alkalmazott statisztikai módszerek alkalmazhatóak-e magyar korpuszra.

2. Az *ádáz* lemma kollokációi két magyar nagykorpuszban

A számítógépen tárolt, szófaj-elemzett és sokszempontú keresésnek alávethető korpuszok megjelenése nagy lehetősége a kollokáció-kutatásnak. Ennek illusztrálására bemutatom az *ádáz* lemma kollokáltjait a két legnagyobb magyar általános korpusz, a 187 millió szövegszavas Magyar Nemzeti Szövegtár[4] (MNSz) és a 25 millió szövegszavas Magyar Történeti Korpusz[5] (MTK) 1944 után keletkezett 6527 szövegének az elemzésével, melyek 10,1 millió szövegszót tartalmaznak.[6] E két korpusz tehát együtt is vizsgálható, bár a korpuszépítés módja és a szövegtípusok eloszlása szempontjából is különböznek.[7]

A korpuszalapú keresés lehetővé teszi, hogy 5-10-15 stb. szavas szövegkörnyezetével együtt kilistázzuk az *ádáz* lemma (*ádáz, ádázabb, legádázabb, ádázul, ádázabbul, ádázan,* stb.) összes előfordulását, és ezen belül arra is rákérdezhessünk, hogy közvetlenül az *ádáz* előtt és/vagy után, vagy 1-2-3 stb. közbeékelődő szóval hányszor fordul elő például az *ellenség* szó vagy például bármely főnév. Általánosabban: a sokszempontú keresés lehetővé teszi, hogy adott szövegszó vagy lemma megadható nagyságú környezetében (*n* szónyi kollokációs ablakon belül) megadható szövegszó vagy lemma vagy megadható szófajú nem specifikált szó összes előfordulását megadható nagyságú szövegkörnyezetben, konkordanciaként listázza ki a program. A MTK 1944 után keletkezett szövegeiben az *ádáz* lemma 87-szer fordul elő; a korpusz ezen részében az *ellenség* lemma 1229-szer, az *ádáz ellenség* kollokáció kilencszer fordul elő (ld. 1. ábra).

1. ábra
Az ádáz *és az* ellenség *lemmák a MTK 1944 utáni állományában*



A MNSz-ban az *ádáz* lemma 578 mondatban 581-szer fordul elő. Az alábbi, 1. és 2. táblázatban gyakorisági sorrendben bemutatom a két korpuszban egynél többször előforduló kollokációit két szófaji csoportban (melléknév + főnév, határozószó + ige/melléknév kollokációk).

1. táblázat
Magyar Történeti Korpusz

MTK (1944-): ***ádáz* (mnév) + (fnév)** n = 68	**MTK (1944-):** ***ádáz* (hatszó) + (ige/mnév)** n = 16
9 ellenség 8 (kenyér)harc 3 ellenfél 3 küzdelem 2 (kártya)csata + 43 hapax	3 figyel 2 gyűlöl + 11 hapax

2. táblázat
Magyar Nemzeti Szövegtár

MNSz: ***ádáz* (mnév) + (fnév)** n = 548	**MNSz:** ***ádáz* (hatszó) + (ige)** n = 30
94 harc 71 ellenség 69 küzdelem 44 csata 39 vita 38 ellenfél 15 verseny 8 háború 8 kampány 7 düh 5 ellenző 4 csatározás 4 ellenállás 4 kutya 3 idő 3 támadás 3 tekintet 2 arckifejezés 2 bíráló 2 erő 2 gyűlölködés 2 harcos 2 kritikus 2 légkör 2 összecsapás 2 párt 2 rivális 2 verekedés 2 vetélkedés 2 vihar + 100 hapax	3 harcol 3 védekezik 2 küzd + 22 hapax

A 3. táblázatban egy mini-felmérés ugyanilyen gyakorisági listái láthatók: a 2002. május 18-án a Nyelvtudományi Intézetben tartott előadásom magyar anyanyelvű nyelvész hallgatóságát kérdeztem meg, az *ádáz* mely kollokáltjai jutnak eszükbe. A 18 önként vállalkozó[8] szabadon választott számú kollokáltat jelölhetett meg írásban. Kontextust nem adtam meg. (Az 1-3. függelékben ABC-sorrendben a két korpuszban talált és a felmérésben felvett összes előfordulást megadom.)

3. táblázat
Felmérés a Nyelvtudományi Intézetben

Felmérés: ***ádáz* (mnév) + (fnév)** n = 37	**Felmérés:** ***ádáz* (hatszó) + (ige)** n = 12
12 ellenség 9 küzdelem 8 harc 5 csata 3 gyűlölet 2 düh 2 vita + 30 hapax	4 küzd 3 harcol + 10 hapax

Látható, hogy a *harc, ellenség/ellenfél, küzdelem* kollokáltak mindegyik listában előkelő helyet szereztek, a *csata, vita* kollokáltak két listában kerültek előre. (A fenti táblázatokból látható, hogy az *ádáz* lemma melléknévként sokkal gyakoribb, mint határozószóként.) A következő fejezetekben az *ádáz* lemma ezen leggyakoribb kollokáltjait elemzem a két korpuszban.

2.1. Az *ádáz* kollokáltjai a Magyar Nemzeti Szövegtárban

A MNSz-ban az *ádáz* lemma 578 mondatban talált 581 előfordulása közül a tíz leggyakoribb, főnévi kollokálttal (*harc, küzdelem, csata, vita, verseny, kampány, háború, düh, ellenség, ellenfél*) 394 esetben fordul együtt elő (ez az *ádáz* előfordulásainak 68%-a, vagyis több, mint kétharmada). 2-5 esetben fordul elő 23 kollokációs párban, egyszer (hapax) 121 kollokációs párban (32%). (A hosszú „hapax-farok" jellemző tulajdonsága minden korpusznak.) Példaként lássuk az *ádáz* + *verseny* 16 előfordulását konkordancialistában! (E kollokáció elemzését ld. a 4.1.6. függelékben.)

press-nar.3042.6.3 egyik bulvártévé, nem kívánván elmaradni az **ádáz** hírversenyben, vagy miben, a másik mögött
press-hvg.4698.5.1 használnak, így érhető, hogy a kibocsátók **ádáz** piaci versenyt vívnak egymással.
press-dtn.1021.6.2 Az **ádáz** válogatóversenyek után örvendetesen szépszámú PVSK-versenyző
press-hir.4588.8.3 **Ádáz** verseny dúl, melyik irodát engedélyezi előbb
press-hir.21625.2.1 **Ádáz** verseny folyik a tankönyvpiacon
press-hir.21625.6.1 mint 7 milliárd forintos piacon zajló **ádáz** verseny pozitív hatásai az elmúlt években
press-hir.26565.4.1 startpisztoly, s egy eddig nem tapasztalt **ádáz** verseny veheti kezdetét, aminek talán

press-hvg.1941.4.1 Kevés szereplővel **ádáz** verseny folyik az építkezések kockázataira köthető
press-nszb.10422.6.1 konferenciaturizmus, nem csoda, hogy e téren **ádáz** verseny folyik az országok és a
press-hir.1182.7.1 mit sem törődve a két- eddig **ádáz** versenyben levő- piac mostantól kézen fogva
press-hvg.526.7.2 10 százalékkal nőtt 1997-ben- az egyre **ádáz**abb versenyben a betéti és a hitelkamatok
press-hvg.763.8.2 egyik meglepő példája, hogy az egymással **ádáz** versenyben lévő cégek a közös érdekek mentén
pers-ind.247751.3.1 televíziók értékei a kereskedelmi televíziókkal vívott **ádáz** versenyben?
press-nszb.15459.10.1 AG. történetéről és a kereskedelmi világcégek **ádáz** versenyéről a Hétvége közöl részletes
press-hir.21943.4.2 A bizonytalan szavazók megnyeréséért folytatott **ádáz** versenyfutása a szocialistákkal még
press-nszv.22583.5.2 kvóták és követelmények súlya alatt kell **ádáz** versenyt folytatniuk a reklámbevételt hozó
press-nszv.34864.8.1 hanem a hitelezési lehetőségekért folytatnak mind **ádáz**abb versenyt egymással.

Az *ádáz* melléknév tíz leggyakoribb főnévi kollokáltjának Magyar Nemzeti Szövegtáron belüli eloszlási mutatóit a 4. táblázat tartalmazza.

4. táblázat

Az ádáz *lemma tíz leggyakoribb kollokációja a MNSz-ban: melléknév + főnév (+ ige)*

ádáz (578*/581*) 'Adj + N' elemeként	közvetlenül#	nem közv.	**össz.**	folyik	dúl	folytat	vív
ádáz + harc (*harc* 16403/16690)	78*	16	**94***	16*	8*	8*	21
ádáz + ellenség (*ellenség* 8533/8874)	68*	3	**71***	-	-	-	-
ádáz + küzdelem (*küzdelem* 9062/9183)	65*	4	**69***	7*	1	18*	11*
ádáz + csata (*csata* 5638/5731)	43*	1	**44***	2	13*	0	13*
ádáz + vita (*vita* 61549/64288)	33*	6	**39***	9	4	3	0
ádáz + ellenfél (*ellenfél* 11719/11827)	30*	8	**38***	-	-	-	-

ádáz (578*/581*) 'Adj + N' elemeként	közvet-lenül#	nem közv.	**össz.**	folyik	dúl	folytat	vív
ádáz + verseny (*verseny* 24857/25534)	15*	1	**16***	3	1	3	2
ádáz + kampány (*kampány* 9762/9941)	7*	1	**8***	1	0	7	0
ádáz + háború (*háború* 23994/24847)	7*	1	**8***	0	2	0	1
ádáz + düh (*düh* 2232/2243)	5*	2	**7***	-	-	-	-
ÖSSZESEN	351	43	**394**				

Jelölések:
n/m: talált mondatok száma/talált szavak száma (tokenek) (A számításokban mégis az első értéket használtam, mivel jelenleg a MNSz keresőprogramja mondatonként csak egy találatot ír ki.)
* A gépi találatszámhoz képest kézzel javított érték.
# A kollokált egyszerű vagy összetett szó.

Nyilvánvaló, hogy ilyen gyakoriságoknál a véletlennél szorosabb kapocs fűzi össze ezen szótöbbesek tagjait. Az alábbiakban először részletesen bemutatok egy ilyen kollokációt az MNSz-ből (2.1.1), majd a MTK hasonló bemutatása következik (2.2-2.2.1). (A 4-5. függelékben a kíváncsi olvasó a további leggyakoribb előfordulásokat is megtekintheti.)

2.1.1 Az *ádáz* (lemma) + *harc* (lemma) kollokáció

Ebben a kollokációban a *harc* önálló lemmaként 84 esetben, összetett szó második tagjaként 10 esetben szerepel (*ádáz konkurenciaharc* (4x), *ádáz pozícióharc* (2x), *ádáz árfolyam-, kultúr-, kenyér-, párharc*).

Nem közvetlenül követik egymást a kollokáltak a következő kollokációkban: *ádáz (elvi) harc; ádáz belső harc; ádáz és látványos harc; ádáz és sok értelmetlen áldozatot követelő kultúrharc; ádáz hatalmi harc; ádáz külső és belső harc; ádáz, nem egyszer fegyveres harc; ádáz piaci harc* (3x)*; ádáz politikai harc* (3x)*; ádáz utcai konkurenciaharc; ádáz utódlási harc*. Állítmányként egy esetben szerepel az *ádáz*: *amely harc éppoly ádáz.* A találati listában további 16, nem azonos frázisban szereplő, tehát érvénytelen találat szerepelt, valamint a *harc* szót az összetétel első tagjaként tartalmazó *ádáz harckészséggel* kifejezés.

A táblázatban szereplő igéken kívül az alábbi igék kapcsolódtak az á*dáz* + *harc* kollokációhoz, alacsony gyakorisággal (egy vagy két esetben):

- ádáz harc kibontakozik (2x)/kirobban (2x)/kitör/indul/lesz/zajlik
- ádáz harcot okoz (2x)/végighallgat/gerjeszt/lát/meghirdet/hirdet/kivált/indít/ígér/elkezd

- ádáz harccal üldöz
- ádáz harcra készül/van kilátás
- ádáz harcba kezd (2x)/keveredik
- ádáz harcban mozog/alulmarad/eldől
- ádáz harcról szól
- ádáz harccá alakul

2.2. Az *ádáz* kollokáltjai a Magyar Történeti Korpuszban

A MTK 1944 után keletkezett szövegeiben az *ádáz* lemma 87 előfordulása közül 84 esetben szerepel melléknévként vagy határozószóként[9], mondat részeként[10]. Az öt leggyakoribb, főnévi kollokálttal (*ellenség, harc, ellenfél, küzdelem, csata*) 25 esetben fordul együtt elő (ez az *ádáz* előfordulásainak 30%-a), mindegyik esetben közvetlenül a főnév előtt. A hapaxok száma (melléknévként ill. határozószóként): 54 (64%). Példaként hadd mutassam be az *ádáz* + *harc* 8 előfordulását konkordancialistában! (E konkordancia elemzését ld. az 5.1.1. alfejezetben.)

1947 BIBÓ ISTVÁN: VÁLASZ és gyakorlatilag azután azt jelentené, hogy **ádáz** egyházellenes harciriadók váltakoznának erőltetett

1957 DÉRY TIBOR: A BEFEJEZETLEN MONDAT hullámzás alatt az ember megsejtette az **ádáz** harcot, amelyet a terjeszkedni

1958 RÓNAY GYÖRGY: PETŐFI ÉS ADY KÖZÖTT álom; abban bontakozhatik ki a lét **ádáz** kenyérharcánál magasabb 220

1959 LOVAS MÁRTON: VALÓSÁG politikusok, vagy akik belevetették magukat a leg**ádáz**abb harcba, s annak, aki

1964 GÁRDOS MARISKA: KUKORICÁN TÉRDEPELVE szociális élet a maga irgalmatlan és **ádáz** kenyérharcával kemény

1976 KIBÉDI VARGA ÁRON: MAGYAR MŰHELY érett (talán csúnya) lány vágyálmait s **ádáz** harcát a

1977 GECSE GUSZTÁV: TÖRTÉNELEM ÉS KERESZTÉNYSÉG akik ellen az ősi gyülekezetek még **ádáz** harcot hirdettek.

1981-1983 RADNÓTI ZSUZSA: Cselekvés-nosztalgia Vagy egy áldozaté, aki a Történelem **ádáz** harcokkal teli

1982 RÓNAY LÁSZLÓ: LITERATURA patronokat"'. A pályakezdő író, aki oly **ádáz** harcokat vívott Osváttal

Az *ádáz* melléknév öt leggyakoribb főnévi kollokáltjának Magyar Történeti Korpuszon belüli eloszlási mutatói az 5. táblázatból olvashatók le.

5. táblázat

Az ádáz *lemma öt leggyakoribb kollokációja a MTK 1944 után keletkezett szövegeiben: melléknév + főnév (+ ige).*

ádáz (84)	közvetlenül[#]	nem közv.	**össz.**	folyik	dúl	folytat	vív
ádáz + ellenség (*ellenség* 1229)	9	0	**9**	-	-	-	-
ádáz + (kenyér)harc (*harc* 2499)	8	0	**8**	0	0	0	1
ádáz + ellenfél (*ellenfél* 521)	3	0	**3**	-	-	-	-
ádáz + küzdelem (*küzdelem* 896)	3	0	**3**	0	0	1	0
ádáz + csata (*csata* 808)	2	0	**2**	0	0	0	1
ÖSSZESEN	25	0	**25**				

[#] A kollokált egyszerű vagy összetett szó.

2.2.1 Az *ádáz* (lemma) + *ellenség* (lemma) kollokáció

A Magyar Történeti Korpuszban az *ellenség* lemma nem szerepel összetett szó elemeként az *ádáz* környezetében. Minden esetben közvetlenül követik egymást a kollokáltak. A találati listában további egy, nem azonos frázisban szereplő, tehát érvénytelen találat szerepelt. („A kétségbeesés s a tehetetlenség *ádáz* dührohama a belső és külső *osztályellenséggel* szemben …" [1964 Gárdos Mariska: Kukoricán térdepelve])

A példák bemutatása után rátérek a kapcsolat szorosságának mérésére, mérhetőségére.

3. Kvantifikálás

3.1 A kollokációs ablak

A számítógépes lexikográfia egyik legnehezebben megoldható problémája az, hogy minél nagyobb arányban gépileg választhassuk ki a **valódi találatokat** az adatjelöltek közül. A szófaj-alapú kódolás (*tagging*) ugyanis többnyire nem ad elegendő információt az egy szerkezetbe tartozó szavakról, a szintaxis-alapú kódolás (*parsing*) pedig a magyar nyelvre egyelőre nem megoldott. Ha nem elégszünk meg a közvetlenül egymás mellett álló szavakon való kereshetőséggel – márpedig miért elégednénk meg ezzel? –, szintaktikai kódolás hiányában

a keresést vagy előre meghatározott ún. szó-ablakon (*word-span*) belül rögzíthetjük; mint fentebb említettem, ennek optimuma az angol nyelvre négy közbeékelődő szó. Kerestethetünk egész mondaton is, ahol a mondatvéget egyszerűen a „mondatvégi írásjel + szóköz + nagybetű" karaktersorozat definiálja. Nézzük meg néhány korpusz-példán, milyen adatjelöltekhez jutunk ezekkel a módszerekkel. Példáimat az *ádáz* + *harc* és az *ádáz* + *düh* kollokációs párok MNSz-beli adatjelöltjeiből választottam. (A vastagon szedett sorszám hibás találatot jelez.)

Közvetlenül egymás mellett:

(1) …egyébként is szerepel, hogy *ádáz harc* dúl az európai nagyvárosok között… (off-fovkozgy.73.141.23)
(2) …épp azokért a kerületekért folyik a *legádázabb harc*, ahol nagyszabású közmunkák… (press-nem.13972.6.3)
(3) …mennyi a sunyi paraszt, mennyi az *ádáz düh* és mennyi a finom epikureizmus… (sci-mek.77.306.16)
(4) …a protestantizmus elleni *harc legádázabb* intézménye… (sci-etu.197.2.3)

Vagyis „melléknév, majd főnév" sorozatra keresve valódi találatokat kapunk. A „főnév, majd melléknév" sorozat hibás találatot hoz (de az ilyen típusú keresést nem szabad kizárni – ld. a 8. példát).

Közvetlenül egymás mellett összetett szóként:

(5) …ugyanakkor *ádáz árfolyamharc* a szubjektumtőzsdén. (sci-mek.77.79.1)

Ha a lehetőség adott, a kollokációkban szereplő összetett szavakat feltétlenül érdemes az utótag lemmájával együtt gyűjteni. Mivel a korpusznyelvészetben a szóköz nagy úr, a nem egybeírt bővítmények szintaktikai elő-elemzés hiányában külön tárgyalandók, még ha e két jelenség hasonló is. (Az összetett szavak vetik fel leginkább a kollokációk produktivitásának kérdését; ld. Sinclair [1997], Lüdeling [2002].)

Egy közbeékelődő szóval:

(6) *Ádáz belső harcokat* okoznak az RMDSZ-ben… (press-nszb.4678.3.1)
(7) …öljük, pusztítjuk egymást, *ádáz, testvéri dühvel*… (lit-dia.7138.1.3.3)
(8) Amely *harc* éppoly *ádáz*, mint amelyből kivált. (lit-dia.11394.19.2)

A fenti példák alapján nyilvánvaló a szó-ablak alapú keresések haszna különböző szerkezetek esetén: míg a 6. és 7. példa egy-egy főnévi csoportot alkot, de még ezek szerkezete is különbözik, addig a 8. példában az *ádáz* melléknév állítmányi helyzetben van. (Szintén állítmányi helyzetben van az *ádáz* melléknév az alábbi esetekben: "Keleti László mint művezető először igen ádáz, a termelés érdekeit

hajszolja." [1979 Haraszti Miklós: Darabbér]; „a feleség pedig minél ádázabb" [1990: Ardai Zoltán: Filmvilág].) Szintaktikailag elemzett korpuszban az adott szerkezeti mintázatok szerinti keresés esetén találnánk meg e példákat, de a közelségen alapuló szó-ablakos keresés is megtalálja őket.

Két közbeékelődő szóval:

(9) *Ádáz* átok, gyilkoló *düh*. (lit-dia.12889.1.5.2)
(10) …a „lázadó" vörösgárditák vívtak *ádáz, nemegyszer fegyveres harcot*… (sci-etu.5835.8.5)
(11) …a voksokért dúló *egyre ádázabb és látványosabb harcban* olyan esetlenül és idegenül mozog… (press-dtn.5931.11.1)

A 9-11. példából kiderül, hogy a szó-ablakos keresés egy több szóból álló összetett szerkezetbe tartozó kollokációt éppúgy adatjelöltként tüntet fel (10-11), mint a közeli, de nem egymáshoz tartozó szópárt (9. példa).

Négynél több közbeékelődő szóval:

(12) Az 1990 és 1994 között vívott *ádáz és sok értelmetlen áldozatot követelkultúrharc*… (press-hir.4170.10.3)
(13) Akár az ördögi véletlennek is tulajdoníthatnánk, hogy ez *ádáz* harckészséggel valóságos csatákat csak polgári háborúkban, belső *harcokban* vívnak. (press-nszb.7787.8.5)
(14) Ezerkilencszáztizenhat tavaszán, a világ *legádázabb* háborújának kellős közepében, mint valaha a treuga deik idejében, egypár órára béke lengi be a *harcok* völgyét. (lit-dia.14296.104.1)
(15) Az újszerűség lehetett az oka, hogy az előkészítés során *ádáz* csatákat kellett vívni: minek nekünk ez a fura nevű, „ombudsman" jelszóval *harcba* indult had, sokan váltig hangoztatták, hogy a gyermekcipőben járó jogállamunkat már annyi intézmény védi, hogy a végén még a sok bába között elvész a gyerek. (sci-mek.368.380.3)

A 12. példa jelzi, hogy a négy közbeékelődő szó szerinti szó-ablakos keresés néha nem elegendő. Ugyanakkor a szó-ablak túlságosan nagyra nyitása rengeteg rossz találatot eredményez (13-15. példa). A mondatablak-alapú keresés adatjelöltként tünteti fel a 13-15. mondatot is. A szintaktikailag elő-elemzett korpuszban ezek, a keresett szavakat különböző szerkezetben tartalmazó esetek nem jelennének meg adatjelöltekként. Sőt, a mondatalapú keresés az alábbiakat is találatként észleli:

(16) A csúcsot a Himalája meghódítóinak segítőiről „serpáknak" nevezett kormányszakértők készítették elő sok hónapos munkával: ezeknek a hamar elfelejtett közleményeknek minden mondatáért *ádáz* harc folyik, amelyet a diplomaták vívnak, ám mindig a politikusok nyernek meg, hiszen a főnökök a hazai porondon használható,

de legalábbis nem ártó megfogalmazásokért küldik *csatába* embereiket. (press-hir.26230.6.5)

(17) Utólag jöttem rá, hogy az ezerszer nagyobb intenzitással átérzett gyermekérzések semmiségeken kitörő *csatáiban* is hányszor milyen fölöslegesen hajszoltam bele magam a düh- és gyűlöletrohamig menő vitákba, mint például egyszer alkonyattájt, az előttünk pompázó nyári kert ihlette kedvenc virágelosztó játékunkban, mikor a nővérem már egyezségre hajló mondatából: De a petúniákat én szeretem jobban, jó? kerekedett az *ádáz* küzdelem, mert úgy éreztem, én ezerszer jobban szeretek mindent. (lit-dia.7648.3.28.16)

A mondatablak-alapú keresésnek ugyanakkor az az előnye, hogy a mondathatár által elválasztott szópárokat nem emeli a találati listába.

Kollokációs szótár összeállításakor a találati lista szűkítése érdekében ehhez a munkálathoz javaslom a fenti keresési módszerek vegyítését: egyrészt a szóalapú kollokációs ablak definiálásakor a fenti karaktersorozat által definiált mondathatárt is vegyük figyelembe, másrészt a statisztikai mutatók kiszámítása előtt a szópárokat engedjük át egy, a szófaji kritérium által definiált szűrőn.

A kollokációs ablak nagyságához az *ádáz*-kollokációk távolságának alábbi eloszlását érdemes figyelembe venni:

- közvetlenül egymás mellett 470 esetben (ebből az *ádáz* lemma a második helyen 2 esetben) – 85%
- egy szótól elválasztva 56 esetben (ebből az *ádáz* lemma a második helyen 1 esetben) – 10%
- két szótól elválasztva 15 esetben (ebből az *ádáz* lemma a második helyen 3 esetben) – 3%
- három szótól elválasztva 4 esetben – 0,7%
- négy szótól elválasztva 1 esetben
- öt szótól elválasztva 2 esetben
- hat szótól elválasztva 6 esetben
- nem elemezhető 28 előfordulás

Ebből az következik, hogy a kollokációs ablakot maximum három közbeékelődő szóra érdemes beállítani. A szavanként haladó gépi keresés mindkét irányban nyitott, de az esetleges kézi kereséseknél az ablakot érdemes úgy beállítani, hogy a vizsgált szó mindkét oldalán keressen. (Például a melléknév állítmányként állhat a hozzá kapcsolódó főnév mögött: *amely harc éppoly ádáz*, stb.).

A következőkben az adatjelöltek kiemeléséhez szükséges kollokációs statisztikai módszereket mutatom be, vagyis azt, hogyan tehetjük mérhetővé a gyakori szótöbbesek tagjainak kapcsolati szorosságát.

3.2. Az asszociáció mérése

A korpuszok általános jellemzésére a leggyakoribb mutatók a korpusz nagysága szövegszószám szerint (***N***), a szövegszavak és a lemmák type/token

eloszlása, a mondatok száma és a szófajeloszlás. Ugyanezeknek a mutatóknak az alkorpuszonként vagy szövegenként különböző eloszlási mintázata szintén gyakran vizsgálat tárgya. A szótöbbesek vizsgálatakor a legalapvetőbb mérőszám a **gyakoriság**, a szótöbbes tagjainak közös gyakorisága (**f(xy)**) a korpuszban; ezt a mutatót használtuk az *ádáz* és kollokáltjai esetében (ld. 4-5. táblázat).

Többet árul el a szótöbbesről, és (al)korpuszközi összehasonlítást is lehetővé tesz a **relatív gyakoriság**, a szótöbbesnek a korpusz összes szövegszavához képest megadott gyakorisága (**f(xy)/N**). Hiszen egy alacsony gyakorisági érték egy kis korpuszban gyakoribb szóra utalhat, mint egy magasabb érték egy nagyobb korpuszban. Hasonlítsuk össze például az *ádáz + ellenség* kollokáció MTK- és MNSz-beli előfordulásait! Bár ez a kollokáció a MNSz-ban 71-szer, a MTK-ban 9-szer fordul elő, a MTK-ban mégis relatíve gyakoribb (f (*ádáz+ellenség*)/N = 0,00000089, mint a MNSz-ban (f (*ádáz+ellenség*)/N = 0,000000473).

3.2.1. A kölcsönös információ együtthatója

A relatív gyakoriság azonban még mindig nem nyújt elegendő segítséget a jó és csakis a jó adatjelöltek korpuszból való kiszűréséhez, ugyanis túl sok esélyt ad a véletlenül bekerülő adatoknak. Nem veszi számításba ugyanis azt, hogy bármelyik kollokált a kollokáción kívül milyen gyakorisággal fordul elő. Két igen gyakori szónál ugyanis akár a „véletlen" (vagyis a szabad asszociáció) műve is lehet, hogy egyszer-egyszer egymás mellé kerülnek. Ha azonban az egyik, vagy mindkét szó az esetek egy bizonyos részében csak a kollokáltja mellett bukkan fel, biztosak lehetünk benne, hogy a két szót a véletlennél szorosabb kapocs, mintegy a kölcsönösség fűzi egymáshoz. Az a mutató, amely a szótöbbes relatív gyakoriságát a szótöbbes két (vagy több) tagjának a kollokáción kívüli relatív gyakoriságának arányában adja meg, a **kölcsönös információ** (pontosabban, az eredmények könnyebb kezelhetősége kedvéért az érték tízes logaritmusával szoktak számolni). Minél magasabb a kölcsönös információ (*mutual information, MI*) értéke, annál szorosabb kapocs fűzi össze a kollokáció tagjait.

$$MI = \log \frac{f(xy)/N}{f(x)/N \ * \ f(y)/N}$$

Más szóval ez a mutató azt jelzi, hogy a kollokáció két tagjának külön-külön előfordulási valószínűsége hogyan viszonyul az együttes előfordulás valószínűségéhez. Vagy egy harmadik, információ-elméleti megfogalmazásban: az érték azt méri, hogy mennyivel csökkenti az egyik szó megjelenése a bizonytalanságot a másik szó megjelenésével kapcsolatban (Manning és Schütze 1999: 171). Minél nagyobb a szám, annál jelentősebb kollokációval állunk szemben, mivel a kollokáció két tagjának előfordulásaihoz képest annál gyakoribb az együttes előfordulás. Például az *ádáz + csata* és az *ádáz + vita* kollokációk majdnem azonos gyakorisággal fordulnak elő a MNSz-ban (44-szer, ill. 39-szer, ld. 4. táblázat). Ugyanakkor míg a *vita* lemma több, mint hatvanezerszer fordul elő

ugyanott (f (*vita*) = 61549), a *csata* szó viszont csak mintegy tizedannyiszor (f (*csata*) = 5638), nyilvánvaló, hogy ez utóbbiból az *ádáz* lemmával való negyven körüli közös előfordulása szorosabb kapcsolatot jelez (MI = 3,3), mint a *vita* lemmával való kapcsolata (MI = 2,2). (Az *ádáz* lemma tíz leggyakoribb kollokáltjának MI-értékeit ld. a 6. táblázatban.)

6. táblázat.

Az ádáz *lemma MNSz-beli tíz leggyakoribb kollokáltjának mutatói*

y	N	f(xy)	f(x=*ádáz*)	f(y)	MI	rangsorrend f (xy) alapján	rangsorrend MI alapján
harc	164000000	94	578	16403	3,17236	1	4
ellenség	164000000	71	578	8533	3,33432	2	1
küzdelem	164000000	69	578	9062	3,29578	3	3
csata	164000000	44	578	5638	3,30649	4	2
vita	164000000	39	578	61549	2,21600	5	8
ellenfél	164000000	38	578	11719	2,92505	6	5
verseny	164000000	16	578	24857	2,19480	7	9
kampány	164000000	8	578	9762	2,32771	8	7
háború	164000000	8	578	23994	1,93715	8	10
düh	164000000	7	578	2232	2,91056	10	6

Hasonlóan magas MI-értékeket kapunk, ha a MTK-ban talált öt leggyakoribb *ádáz*-kollokációt vizsgáljuk meg.

7. táblázat

Az ádáz *lemma öt leggyakoribb kollokáltjának mutatói a MTK 1944 utáni szövegeiben*

y	N	f(xy)	f(x=*ádáz*)	f(y)	MI	rangsorrend f (xy) alapján	rangsorrend MI alapján
ellenség	10100000	9	84	1229	2,944733	1	1
harc	10100000	8	84	2499	2,585366	2	4
ellenfél	10100000	3	84	521	2,840326	3	2
küzdelem	10100000	3	84	896	2,604855	3	3
csata	10100000	2	84	808	2,473661	5	5

A kölcsönös információ páronkénti/többesenkénti kiszámításával sorba rendezhetjük a korpusz szópárjait/szótöbbeseit. A fenti két táblázatból láthatjuk, hogy a kollokációs adatjelölteknek ez a rangsora, vagyis az ún. **szignifikancia-lista** *(significance list)* az MI-mutató és a gyakoriság figyelembevételével nem lenne azonos. A szignifikancia-listán egy alkalmas küszöbérték meghúzásával megkapjuk azon adatjelöltek listáját, melyek a kézi elemzés alapját jelentik.

A kölcsönös információ mutatójának segítségével pontos képet kapunk arról is, hogy egyes szinonimák szabadon kombinálódnak-e, vagy használatuk komplementáris. A 8. táblázatban a *sovány-vékony* szinonimapár segítségével ezt mutatom be. A korpuszok azt mutatják, hogy az embereknek inkább *sovány arca* van, mint *vékony arca*. A *férfiak* inkább *sovány*ak, a *nők* inkább *vékony*ak. A komplementáris eloszlásoknál (a 8. táblázatban szürkével jelölve) feltétlenül kollokációkkal van dolgunk. A *vékony hang* többször (57-szer) szerepel, mint *vékony sugár* (21), de mivel a *hang* lemma gyakorisága sokkal nagyobb (f (*hang*) = 29194), mint a *sugár* lemmáé (f (*sugár*) = 2368), ezért a gyakoribb *hang* előforduláshoz képest a *vékony hang* együttes előfordulása alacsonyabb értéket ér el (MI (*vékony hang*) = 2,1), mint a *vékony sugár*é (MI (*vékony sugár*) = 3,4).[11] Az MI-t használja Sass (2007) is az igék és vonzataik kollokációs szorosságának mérésére a „Mazsola”-programban (http://corpus.nytud.hu/mazsola/).

8. táblázat

A sovány *és* vékony *szinonimák néhány kollokáltja (MTK 1944-, MNSz)*

	sovány (MTK+MNSz)	MI(MNSz)	*vékony* (MTK+MNSz)	MI (MNSz)
vigasz	60 (11 + 49)	4,7	1 (1 + 0)	-
férfi	30 (13 + 17)	2,6	14 (10 + 4)	1,5
eredmény	11 (1 + 10)	2,2	-	-
arc	34 (27 + 7)	2,1	11 (8 + 3)	1,7
hús	3 (0 + 3)	1,6	-	-
teremtés	2 (2 + 0)	-	2 (2 + 0)	-
testalkat	-	-	-	-
derék	1 (1 + 0)	-	11 (5 + 6)	1,8
nő	4 (1 + 3)	1,9	17 (4 + 13)	2,1
hang	-	-	57 (32 + 25)	2,1
sugár	-	-	21 (2 + 19)	3,4
réteg	-	-	52 (20 + 32)	3,5
szelet	-	-	95 (6 + 89)	4,0

Az MI hátulütője, hogy az alacsony gyakoriságú szavak, a hapaxok esetében túl magas értéket ad. Ugyanis ha két hapax egymás mellé kerül egy korpuszban (és ez még nagy korpuszok esetén is igen gyakori), a két szó előfordulásaihoz képest az együttes előfordulás maximális lesz (mivel a két szó csak együtt fordul elő). Az alacsony gyakoriságú szavakból álló szópárok magasabb MI-értéket fognak kapni, mint a magas gyakoriságú szavakból álló szópárok, ami ellentmond annak az intuíciónknak, hogy a több adattal alátámasztott mutató erősebb bizonyítékot jelent (Manning és Schütze (1999: 170)).

9. táblázat

Az ádáz *lemma néhány hapax-kollokáltjának gyakorisági és MI-mutatói (MNSz, MTK 1944-)*

	y	N	f(xy)	f(x=*ádáz*)	f(y)	MI
M N Sz	*arc*	164000000	1	578	26414	0,992329
	harcol	164000000	1	578	3904	1,822654
	ibolyaillat	164000000	1	578	8	4,511073
	kritikus (N)	164000000	1	578	2386	2,036493
	kulcscsomócsörgés	164000000	1	578	1	5,414163
	támadássorozat	164000000	1	578	87	3,474644
M T K	*arc*	10100000	1	84	680	2,247533
	Erynnis/Erinnys	10100000	1	84	3	4,602921
	nép	10100000	1	84	7411	1,210165
	uszítás	10100000	1	84	34	3,548563

A 9. táblázatból leolvasható, hogy ez valóban így van. Csak magas gyakoriságú kollokált (pl. f (*arc*) = 26414 a MNSz-ban) esetében kapunk viszonylag alacsony MI értéket (MI (*ádáz arc*) = 0,99 a MNSz-ban) egy olyan szótöbbes esetében, melyet anyanyelvi intuíciónk alapján sem érzünk összetartozónak. Ha a kollokált gyakorisága is alacsony (pl. *ibolyaillat, kulcscsomócsörgés* a MNSz-ban), az érték könnyen MI = 4 fölé szökhet össze nem tartozónak érzett kifejezések esetében is (*ádáz ibolyaillat, ádáz kulcscsomócsörgés*). Más a helyzet az *ádáz Erinnys* szótöbbessel, azt e sorok írója kollokációnak érzi (ld. Berzsenyi: A magyarokhoz) – de ez a probléma már a szótöbbesek kvantifikálhatóságán kívül, e tanulmány elején érintett szemantikai döntések körébe esik.

Ha a függőség mérésére az MI ezért nem is megbízható mutató, a függetlenséget megbízhatóan méri, ezért azzal a megszorítással érdemes használni, hogy csak a három vagy annál gyakoribb szavakat (f (x) > 2 ÉS f (y) > 2) vonjuk be a számításba és az összehasonlításba (Manning és Schütze 1999: 170).[12]

3.2.2. A khinégyzet-próba

A statisztikából jól ismert khinégyzet-próbát itt a szótöbbes tagjai függetlenségének vizsgálatára alkalmazzuk. Ez a próba azt méri, hogy ha – jelen esetben – két változó (pl. *ádáz* és *harc*) függetlenek lennének egymástól, milyen gyakorisággal kerülnének – ezek szerint véletlenül – egymás mellé az adott nagyságú korpuszban; ez az adott változó *várható értéke* (E). Ha a várható értékek jelentősen, szignifikánsan eltérnek a kapott értékektől, akkor elvethetjük azt a feltételezést, hogy a két változó független egymástól, vagyis kicsi a kockázata annak, hogy tévesen tételezzük fel, hogy összefüggenek egymással. Az alábbiakban az elemzés alapját képező kereszttáblában mutatom be az *ádáz – harc* összefüggését a MNSz eloszlásai alapján (10. táblázat).

10. táblázat

Az ádáz *és a* harc *lemmák gyakorisági eloszlása a MNSz-ban*[13]

MNSz	*harc*	nem *harc*	
ádáz	94 *ádáz harc*	484 pl. *ádáz verseny*	578
nem *ádáz*	16309 pl. *véres harc*	163983113 pl. *szelíd őzike*	163999422
	16403	163983597	164000000

Például az *ádáz* + *harc* cella várható értéke E = 0,06, ami azt jelenti, hogy ennyi esélyük lenne véletlenül egymás mellé kerülni egy 164 millió szövegszavas korpuszban, más szóval, csak egy 17-szer nagyobb, 2,5 milliárd szövegszavas korpuszban fordulnának biztosan elő egymás mellett. Mivel 94-szer kerültek egymás mellé, ez csak igen kis valószínűséggel ($p < 0{,}001$) lehet a véletlen műve.

A 11. táblázatból leolvashatók a MNSz és a MTK tíz illetve öt leggyakoribb *ádáz*-kollokáltjainak, valamint a feljebb is vizsgált néhány hapax-kollokáltjának khinégyzet-értékei. A khinégyzet-próba azonban a MI-mutatóhoz hasonlóan a ritkán előforduló vagy hapax-lemmát (pl. *ibolyaillat, kulcscsomócsörgés*) is tartalmazó kollokációk itt is magas értéket fognak adni. Megoldásként a mutató várható értékkel való súlyozását ($khi^2 * E$) javaslom, így ugyanis ellensúlyozzuk az alacsony gyakoriság és az alacsony várható érték miatti magas khinégyzet-értékeket. A 11. táblázat jobb szélső oszlopa ezt az értéket mutatja. A szignifikancia-szint automatikusan[14] megadja azt a küszöbértéket, amely fölött az adott szótöbbest kollokációs adatjelöltnek kell tekintenünk; így az *ádáz harc, ádáz ellenség, ádáz küzdelem, ádáz csata, ádáz vita, ádáz ellenfél, ádáz verseny, ádáz háború, ádáz kampány, ádáz düh* bekerül a lexikográfusok elé kerülő adatjelölt-listába, az *ádáz ibolyaillat, ádáz kulcscsomócsörgés, ádáz támadássorozat, ádáz kritikus, ádázul harcol, ádáz arc, ádáz*

Erinnys, áldz uszítás, áldz nép szókapcsolat viszont nem, ami nagyjából megfelel anyanyelvi intuíciónknak.

11. táblázat

Az áldz *lemma leggyakoribb és néhány hapax-kollokáltjának gyakorisági és khinégyzet-mutatói (MNSz, MTK 1944-).*
**** p < 0,001, ** p < 0,01, * p < 0,05*[15]

	y	f (xy)	N	f (x)	f (y)	E	khi^2	khi^2 x E
M N Sz	harc	94	164000000	578	16403	.063206	139624.22	8825.12***
	ellenség	71	164000000	578	8533	.032880	153180.12	5036.64***
	küzdelem	69	164000000	578	9062	.034919	136215.26	4756.49***
	csata	44	164000000	578	5638	.021725	89029.25	1934.17***
	vita	39	164000000	578	61549	.237169	6338.02	1503.18***
	ellenfél	38	164000000	578	11719	.045157	31903.83	1440.69***
	verseny	15	164000000	578	24857	.095782	2319.57	222.17***
	háború	8	164000000	578	9762	.037616	1685.55	63.40***
	kampány	8	164000000	578	23994	.092457	676.42	62.54***
	düh	7	164000000	578	2232	.008601	5683.37	48.88***
	ibolyaillat	1	164000000	578	8	.000031	32437.57	1.00
	kulcscsomó-csörgés	1	164000000	578	1	.000004	259514.57	1.00
	támadássorozat	1	164000000	578	87	.000335	2980.95	1.00
	kritikus (N)	1	164000000	578	2386	.009194	106.78	.98
	harcol	1	164000000	578	3904	.015043	64.49	.97
	arc	1	164000000	578	26414	.101782	7.93	.81
M T K	ellenség	9	10100000	84	1229	.010221	7907.60	80.83***
	harc	8	10100000	84	2499	.020784	3064.13	63.68***
	ellenfél	3	10100000	84	521	.004333	2071.18	8.97**
	küzdelem	3	10100000	84	896	.007452	1201.87	8.96**
	csata	2	10100000	84	808	.006720	591.30	3.97*
	Erynnis/Erinnys	1	10100000	84	3	.000025	40077.71	1.00
	uszítás	1	10100000	84	34	.000283	3534.46	1.00
	arc	1	10100000	84	680	.005655	174.84	.99
	nép	1	10100000	84	7411	.061636	14.30	.88

Az alacsony gyakoriságú adatjelölteket valójában érdemes még a gépi számítások előtt kivonni az adatjelölt-listából, még akkor is, ha ezzel adatjelöltjeink nagyobb részétől búcsút kell vennünk. Nemcsak a statisztikai mutatók működnek rosszul ezekben az esetekben; a hapax-adatjelöltek között eleve kevés a valódi kollokáció. Evert és Krenn (2001: 6-7) ezt kézi és statisztikai módszerrel is bebizonyította. Az asszociációs mércék fenti bemutatásakor kizárólag ennek alátámasztására vontam be a hapaxokat. A hapax-előfordulások közti valódi kollokációk alacsony aránya összefügg azzal is, hogy a szótöbbesek definíciójában eleve kizártam a ritka előfordulásokat.

A fentiekben kollokációs kutatásokban alkalmazott asszociációs mércéket teszteltem a MNSz és a MTK egy lemmájához kötődő leggyakoribb és legritkább szótöbbesein a célból, hogy a számítások és nyelvi intuíciónk összevetésével ezeken bizonyítsam, e számítások nagy tömegű adaton is alkalmazhatók. A számításokról bebizonyítottam, hogy bizonyos megszorításokkal a leggyakoribb és legritkább adatokon megbízhatóan működnek. Hogy teljes korpuszon való futtatásuk esetén is olyan szignifikancia-listákat eredményeznek, amelyek mind pontosságukban (*precision*), mind lefedettségükben (*recall*) az elvárható hibaarány alatt maradnak, az a próbafuttatások eredményeiből fog kiderülni.

A fentiekben bemutatott gyakoriságon, kölcsönös információn és khinégyzet-próbán túl a kollokációs asszociáció mérésére többek között a t-próbát (ill. a z-próbát)[16], a log-likelihood-arányt és a faktorelemzést szokták alkalmazni; ezek bemutatásával hely hiányában nem foglalkozom. A kollokációk statisztikai alapú kigyűjtéséhez – részben a többi alkalmazott statisztikai mutató hátrányai miatt – mindenképpen valamelyik, fent részletesen ismertetett mutatót javaslom alkalmazni. A kérdés, hogy melyik mutató a legmegbízhatóbb, meggyőzően csak a próba futtatások eredményeinek ismeretében válaszolható meg.[17]

5. Összegzés

Nyilvánvaló, hogy sem a lexikográfiában, sem a szemantikában nem kerülhető el a szónál nagyobb egységek elemzésének komplexitásával való szembenézés. A firthi hagyományból[18] kiindulva Sinclair (1997) *The lexical item* című munkájában fejti ki, hogy a szó mint lexikai tétel nem adekvát, mert adott szó szemantikai környezetének struktúrájában értelmezendő, abban nyeri el jelentését. "Egy szöveg bármely pontján egy adat előfordulását annak függvényében értelmezhetjük, hogy adott környezetben milyen más lehetőségek álltak fenn még. Így minden adat egyrészt saját jogán adat, másrészt más adatok környezetének egyik komponense." (1997: 6) A jelentést a paradigmatikus jelentések közötti választás hozza létre, a szintagmatikus környezet pedig behatárolja a komplex jelentésösszességet, s a két dimenzió egyensúlyban van.

A szöveg jelentését tehát olyan modellel kell leírni, amely minden egyes választási helyen számbaveszi a választás paradigmatikus és szintagmatikus dimenzióit. A fent bemutatott kollokáció-kutatási irány éppen ezen gondolatmenet mentén tartja fontosnak a szótöbbesek korpusz-alapú elemzését.

A továbbiakban a Magyar Történeti Korpuszból gépi úton kinyert, az egyelőre szófaji szűrőn túljutó adatok statisztikai elemzésével olyan adatjelölt-listák kézi elemzése a feladat, amelyek egyre nagyobb arányban tartalmaznak valódi találatokat és a Magyar Nemzeti Szövegtárból.

Irodalom

Benson, M., Benson, E., Ilson, R. 1986. *The BBI dictionary of English word combinations*. Amsterdam: Benjamins.

Evert, S., Krenn, B. 2001. Methods for the qualitative evaluation of lexical association measures. www.ims.uni-stuttgart.de/projekte/corplex/paper/evert/Evert-Krenn2001.pdf. Letöltve: 2001. szeptember 15.

Firth, J. R. 1957. A Synopsis of Linguistic Theory 1930–1955. In: *Studies in Linguistic Analysis*. Oxford: Philological Society; reprinted in Palmer, F. (szerk. 1968) *Selected Papers of J. R. Firth*. Harlow: Longman.

Heid, U. 2002. Collocations in lexicography. A C*omputational Approaches to Collocations* c. worshopon tartott előadás kézirata.

Ittzés N. 2002. *A Nagyszótár szerkesztési szabályzata*. Kézirat. Budapest: MTA Nyelvtudományi Intézet.

Krenn, B., Evert, S. 2001. Can we do better than frequency? A case study on extracting PP[prepositional phrase]-verb collocations. In: *Proceedings of the Association for Computational Linguistics workshop on collocations.* Toulouse. Ugyanez a www.ims.uni-stuttgart.de/projekte/corplex/paper/evert/ KrennEvert2001.pdf cím alatt is.

Manning, Ch. D., Schütze, H. 1999. Chapter 5: Collocations. In: *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.

O. Nagy G. 1966/1982. *Magyar szólások és közmondások*. Budapest: Akadémiai Kiadó.

Pajzs J. 2000. Frazeológiai egységek a Nagyszótárban. In: Gecső T. (szerk.) *Lexikális jelentés, aktuális jelentés.* Budapest: Tinta. 217–226.

Sass B. 2007. „Mazsola" – eszköz a magyar igék bővítményszerkezetének vizsgálatára. In: Váradi T. (szerk.) *Válogatás az I. Alkalmazott Nyelvészeti Doktorandusz Konferencia előadásaiból,* Budapest: MTA Nyelvtudományi Intézet, 117–129. http://www.nytud.hu/alknyelvdok07/proceedings07/Sass.pdf. Letöltve: 2010. november 1.

Sinclair, J. 1991. *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Sinclair, J. 1997. *The lexical item.* The Tuscan Word Centre.

Van der Meer, G. 1998. Collocations as one particular type of conventional word combination. Their definition and character. *EURALEX '98 Proceedings.* Liege: University of Liège, 313–322.

Váradi T. 2002. The Hungarian National Corpus. In: *Proceedings of the 3rd LREC Conference,* Las Palmas, Spanyolország, 385–389, http://corpus.nytud.hu/mnsz

Jegyezetek

[1] "Állandósult szókapcsolatoknak, frazeológiai egységeknek tekintjük azokat a kötött szerkezeteket, amelyekben az egyes elemek részben vagy teljesen elvesztik önálló jelentéstartalmukat, új lexémát hoznak létre, új jelentéssel."

[2] Állandósult szókapcsolatnak nevezi „a több szóból álló, de lexémaszerűen összeforrt, lexémaszerűen viselkedő, külön értelmezésre szoruló nyelvi elemeket", s ezen

belül "é*rtelmezett szókapcsolatnak* az olyan lexémaszerűen összeforrt szintagmát tekint[i], amelynek elemei – de legalábbis egyik elemük – még őrzik eredeti konkrét jelentésüket. Az értelmezett szókapcsolat lehet jelzős, illetve igéből és névszóból álló szószerkezet, ritkábban két névszó határozós v. határozói szerepű szerkezete (pl. fok- v. mértékhatározós szerkezet, kettős határozók)."

[3] A kollokációkat két csoportban tárgyalja: lexikai kollokációkként, melyek főnévből és/vagy igéből és/vagy melléknévből és/vagy határozószóból álló csoportot képeznek, míg a grammatikai kollokációk egy domináns szóból (főnév, melléknév, ige) és a hozzá kötődő prepozícióból vagy infinitívuszos szerkezetből vagy mellékmondatból. (Ez utóbbi szerkezeteket a kollokációs irodalom nagyrészt nem tárgyalja.)

[4] http://corpus.nytud.hu/mnsz/ A magyarországi korpusz 164,7 millió, a határon túli (szlovákiai, kárpátaljai, erdélyi, vajdasági) korpusz 22,9 millió szövegszavas, mindkettő öt alkorpuszban (sajtó 44%, szépirodalmi 22%, tudományos 11%, hivatali 11%, személyes közlés 11%) (Váradi 2002).

[5] www.nytud.hu/hhc. 1772-1992 között keletkezett szövegek (szépirodalom 31%, egyéb próza 51%, vers 9%, dráma 6%)

[6] Köszönöm Pajzs Júliának és Váradi Tamásnak, hogy a Magyar Történeti Korpuszhoz és a Magyar Nemzeti Szövegtárhoz hozzáférhettem. Itt ragadom meg az alkalmat, hogy Pajzs Júliának megköszönjem a szakirodalmi és személyes segítségét is.

[7] Csak minimális átfedést tapasztaltam a szövegminták között. (Esterházy Péter: *A fuharosok* című szövege például mindkét korpuszban megtalálható.) A duplumok kiszűrésére érdemes lenne összevetni elsősorban a MTK 1944 után keletkezett szövegeit és a MNSz szépirodalmi alkorpuszát.

[8] Hadd köszönjem meg ez úton is a segítségüket.

[9] Egy esetben főnévként szerepel: „másfelől éktelen ádázatok undokitanak" (1981 Kerényi Ferenc: A régi magyar színpadon)

[10] Két találat önálló mondatot képez: „Láttam nővéreim rettenetes arcát! – ádáz! ádáz! – Drágáim, az istenért, dehát mi történt?" (1983 Esterházy Péter: *Fuharosok*)

[11] Gépi kvantifikáláskor egyelőre nem tudjuk kezelni a szemantikai kérdéseket, így a kollokáltak eltérő jelentéseit sem (pl. *sugár*$_{1,2}$, *teremtés*$_{1,2}$). Ettől a nehézségtől a fenti számítások során eltekintettem.

[12] A 8. táblázatban ezért nem tüntettem fel az ennél kevésbé gyakori szavak MI-értékeit.

[13] A khinégyzet-eloszlás kétszer kettes táblánál egy egyszerűsített egyenlettel számolható ki:

$$\chi^2 = \frac{N\,(f_{11}\,f_{22} - f_{12}\,f_{21})^2}{(f_{11} + f_{12})\,(f_{11} + f_{21})\,(f_{12} + f_{22})\,(f_{21} + f_{22})}$$

[14] Az MI-mutatóval ellentétben, ahol a küszöbértéket *ad hoc* húzzuk meg.

[15] Szabadságfok = 1-nél a khinégyzet-eloszlás küszöbértékei: p = 0,05 valószínűségi szintnél χ^2 = 3,84; p = 0,01-nél χ^2 = 6,64; p = 0,001-nél χ^2 = 10,8.

[16] E két próbának a kollokációkutatásban alapvető hátránya, hogy normális eloszlású valószínűségeket feltételez, mely feltételnek a lexikográfiai adatok nem felelnek meg. Ellenben Krenn és Evert (2001) éppen a t-próba megbízhatóságát mutatja ki, ld. a következő lábjegyzetben.

[17] Krenn és Evert (2001) hét statisztikai próbát, köztük a MI-t és a khi-négyzet-próbát vetette össze két német nyelvű korpusz, egy nyolcmillió szövegszavas újságnyelvi és egy tízmillió szövegszavas newsgroup-korpusz alapján a fenti kérdés

megválaszolására. Ők arra a meglepő megállapításra jutottak, hogy a „szimpla" közös gyakoriság (f (xy)) alapján rangsorolt szignifikancia-lista az összes asszociációs együtthatónál megbízhatóbb eredményt produkált az általuk vizsgált, két előfordulásnál gyakoribb 'vonzatos ige + elöljárós szerkezet' csoportban (n = 10396). Egyedül az MI-mutató ért el a közös gyakoriságnál szignifikánsan jobb eredményt, de csak akkor, ha kizárólag az MI-mutató 4,0 – 7,5 közötti értékeit elérő szókapcsolatokat vették figyelembe. Evert és Krenn (2001) pedig azt mutatja ki, hogy a német melléknév + főnév kollokációk mérésére a log-likelihood- és a t-próba, az ige + prepozíciós főnév kollokációkéra a t-próba és a gyakoriság a legmegbízhatóbb mérce mind a pontosság, mind a lefedettség dimenziójában.

[18] Firth (1957): „you shall know a word by the company it keeps", vagyis madarat tolláról, szót a környezetéről lehet jobban megismerni.

1. függelék

Az *ádáz* lemmát tartalmazó kifejezések és gyakoriságuk a MNSz-ban.
A szám nélkül feltüntetett kifejezések egyszer fordulnak elő.

mnév + főnév	értelmiségi	intézmény	népharag
acsarkodás	érzés	8 kampány	név
akciókrimi	fegyver	kardtusa	2 összecsapás
aknázás	fenekedés	katona	összetűzés
arc	fenyegetés	katonasors	palástolás
2 arckifejezés	figyelem	kedv	panasz
asszimiláció	filmpárbaj	kegyetlenség	paraszt
állapot	forma	kéj	2 párt
áralku	fulánk	kény	pergőtűz
átok	gazember	kép	pillanat
barát	gond	kéz	1 propaganda (+ 1
birkózás	gyűlölet	kiszorítósdi	kormánypropaganda)
2 bíráló	2 gyűlölködés	klímafront	rafinéria
bűn	8 háború	konkurencia	rakétavetés
célszerűség	1 hadjárat	konkurens	rivalizálás
44 csata	(+ 1	kötélhúzás	2 rivális
4 csatározás	rágalomhadjárat)	kőrözés	romboló
dadogás	hang	2 kritikus	rovarciripelés
dilemma	hangvétel	kulcscsomócsörgés	sereg
dummerauguszt	harag	kupeckedés	számítástechnikus
7 düh	94 harc	4 kutya	szándék
ekletikus(N)	harckészség	(*Ádáz* nevű kutya)	szem
4 ellenállás	2 harcos	69 küzdelem	szenvedély
38 ellenfél	háborúság	2 légkör	szó-birok
71 ellenség	ibolyaillat	magyar(N)	3 támadás
ellenszenv	ideológiakritika	marakodás	támadássorozat
ellenzék	3 idő	mód	3 tekintet
5 ellenző	időszak	nagymenő	településfejlesztés
2 erő	igazságtalanság	nagynéni	tevékenység
erődrendszer	indulat	negyvennyolcas(N)	téma

tudás
tusa
1 tűz (+ 1 ágyútűz)
vekker
2 verekedés
vers
15 verseny
versengés
veszekedés
2 vetélkedés
védangyal
viadal
2 vihar
világ
1 viszály
(+1 csoportviszály)
39 vita

hatszó + ige
bontogat
csatázik
előregördít
fenekedik
figyel
gúnyol
gyűlöl
hadakozik
hadonászik
3 harcol
herreg
kampányol
kardoskodik
kritizál
2 küzd
marakodik
megvillan
nevet
örkődik
ragaszkodik
rávillog
támad
3 védekezik
viszonoz
vív

2. függelék

Az *ádáz* lemmát tartalmazó kifejezések és gyakoriságuk a MTK 1944 után keletkezett szövegeiben.

A szám nélkül feltüntetett kifejezések egyszer fordulnak elő.

mnév + főnév
arc
arckifejezés
Artemón
bosszú
csata + kártyacsata
dühroham
ellenállás
3 ellenfél
9 ellenség
Erinnys
éjszaka
érdek
feleség
fényűzés
fogcsikorgás
gyűlölet
gyűlölködés
harag
6 harc +2 kenyérharc
harciriadó
kamatháború
katona
káprázat
kentaur
király
kulák
3 küzdelem
legenda
művezető
nap
nem (N)
nép
örvény
perc
potenciálkülönbség
Próteüsz
sárkány
szakasz
szellem
szomorúság
szülő
taglejtés
talp
támadás
támadó
uszítás
vágyakozás
vita

hatszó + ige/mnév
3 figyel
2 gyűlöl
hallgat
hisz
iszik
jön
ketyeg
körbevizslat
körülnéz
ráveti magát
utál
üldöz
igényes

3. függelék

Nyelvészek által gyűjtött, az *ádáz* lemmát tartalmazó kifejezések (a 2002. május 18-án a Nyelvtudományi Intézetben tartott előadásom hallgatóságából 18 adatközlő).
A szám nélkül feltüntetett kifejezések egyszer fordultak elő.

mnév + főnév	8 harc	sereg	**hatszó + ige**
áruló	harcos (N)	sor	egymásnak esik
betegség	9 küzdelem	szomszéd	egymást gyötri
bosszú	kölyök	támadás	ellenáll
5 csata	kutya	tekintet	gyűlölködik
2 düh	küzdelem	tolvaj	3 harcol
ellenállás	merénylet	török	4 küzd
ellenfél	mostoha	viadal	néz
12 ellenség	nézés	vihar	ráront
gondolat	nőszemély	viszály + viszálykodás	rátámad
gyilkos	összecsapás	veszekedés	rátör
3 gyűlölet	pillantás	2 vita	támad
hajsza	rabló		viselkedik

4. függelék

Az *ádáz* további leggyakoribb kollokáltjai a Magyar Nemzeti Szövegtárban

4.1.1 Az *ádáz* (lemma) + *ellenség* (lemma) kollokáció

Az *ellenség* lemma nem szerepel összetett szó elemeként az *ádáz* környezetében.

Nem közvetlenül követik egymást a kollokáltak a következő kollokációkban: *ádáz (ám intelligens) ellenség; ádáz belföldi ellenség; ádáz és gyűlölt ellenség.* A találati listában további hat, nem azonos frázisban szereplő, tehát érvénytelen találat szerepelt.

4.1.2 Az *ádáz* (lemma) + *küzdelem* (lemma) kollokáció

Ebben a kollokációban a *küzdelem* lemma nem szerepel összetett szó elemeként az *ádáz* környezetében.

Nem közvetlenül követik egymást a kollokáltak a következő kollokációkban: *ádáz és terméketlen küzdelem; ádáz felekezeti küzdelem; ádáz politikai küzdelem; ádáz választási küzdelem.* A találati listában további öt, nem azonos frázisban szereplő, tehát érvénytelen találat szerepelt.

A táblázatban szereplő igéken kívül az alábbi igék kapcsolódtak az á*dáz* + *küzdelem* kollokációhoz, alacsony gyakorisággal:

- *ádáz küzdelem bizonyítja/indul/vmiről szól /megvisel /kerekedik /kibontakozik /beárnyékol/tart/várható*
- *ádáz küzdelmet bevilágít/emleget/figyel/figyelemmel kísér/felelevenít*
- *ádáz küzdelembe bonyolódik*
- *ádáz küzdelemben áll*

4.1.3 Az *ádáz* (lemma) + *csata* (lemma) kollokáció

Ebben a kollokációban a *csata* önálló lemmaként 40 esetben, összetett szó második tagjaként 4 esetben szerepel (*ádáz sprintcsata* (2x), *ádáz szócsata, ádáz szócsata*).

Nem közvetlenül követik egymást a kollokáltak a *ádáz politikai csata* esetében. A találati listában további három, nem azonos frázisban szereplő, tehát érvénytelen találat szerepelt.

A táblázatban szereplő igéken kívül az alábbi igék kapcsolódtak az á*dáz + csata* kollokációhoz, alacsony gyakorisággal:

- *ádáz csata folytatódik/zajlik/várható/agyongyötör/kitör*
- *ádáz csatát hoz*
- *ádáz csatára vezet*
- *ádáz csatában részt vesz*

4.1.4 Az *ádáz* (lemma) + *vita* (lemma) kollokáció

Ebben a kollokációban a *vita* lemma nem szerepel összetett szó elemeként az *ádáz* környezetében.

Nem közvetlenül követik egymást a kollokáltak a következő kollokációkban: *ádáz belpolitikai vita; ádáz generációs vita; ádáz költségvetési vita; ádáz, már-már botránnyal, kormányválsággal fenyegető vita; ádáz, párton belüli vita; ádáz területi vita.* A találati listában további kettő, nem azonos frázisban szereplő, tehát érvénytelen találat szerepelt.

A táblázatban szereplő igéken kívül az alábbi igék kapcsolódtak az á*dáz + vita* kollokációhoz, alacsony gyakorisággal:

- *ádáz vita kialakul/szól vmiről/nyúlik (hosszúra)/kibontakozik/indul/megindul/ kirobban/bevezet vmit/belül marad vmin/zajlik*
- *ádáz vitát kivált*
- *ádáz vitával tölt (időt)*

4.1.5 Az *ádáz* (lemma) + *ellenfél* (lemma) kollokáció

Az *ellenfél* lemma nem szerepel összetett szó elemeként az *ádáz* környezetében.

Nem közvetlenül követik egymást a kollokáltak a következő kollokációkban: *ádáz külföldi ellenfél; ádáz, felkészült ellenfél; ádáz politikai ellenfél* (4x)*; ádáz politikai bizottsági ellenfél; ádáz republikánus ellenfél.* A találati listában további négy, nem azonos frázisban szereplő, tehát érvénytelen találat szerepelt.

4.1.6 Az *ádáz* (lemma) + *verseny* (lemma) kollokáció

Ebben a kollokációban a *verseny* önálló lemmaként 14, összetett szó második tagjaként két esetben szerepel (*ádáz hírverseny, ádáz válogatóverseny*).

Nem közvetlenül követik egymást a kollokáltak az *ádáz piaci verseny* esetében. A találati listában további négy, nem azonos frázisban szereplő, tehát érvénytelen találat szerepelt (így – példaként – a fenti táblázatban: *ádáz verseny-futás*).

A táblázatban szereplő igéken kívül az alábbi igék kapcsolódtak az á*dáz + verseny* kollokációhoz, alacsony gyakorisággal:

- *ádáz verseny kezdetét veszi/ zajlik*
- *ádáz versenyben elmarad/van* (2x)
- *ádáz versenyről közöl vmit*

4.1.7 Az *ádáz* (lemma) + *kampány* (lemma) kollokáció

Ebben a kollokációban a *kampány* lemma nem szerepel összetett szó elemeként az *ádáz* környezetében.

Nem közvetlenül követik egymást a kollokáltak az *ádáz választási kampány* kollokációkban. A találati listában további két, nem azonos frázisban szereplő, tehát érvénytelen találat szerepelt.

Az ádáz + kampány kollokációhoz a táblázatban feltüntetett *folyik, folytat* igéken kívül más ige nem kapcsolódott.

4.1.8 Az *ádáz* (lemma) + *háború* (lemma) kollokáció

Ebben a kollokációban a *háború* önálló lemmaként öt, összetett szó második tagjaként két esetben szerepel (*ádáz vallásháború, ádáz polgárháború*).

Nem a melléknév-főnév sorrendben követik egymást a kollokáltak a *háborúja ádáz volt* kifejezésben. A találati listában további 12, nem azonos frázisban szereplő, tehát érvénytelen találat szerepelt.

A táblázatban szereplő igéken kívül az alábbi igék kapcsolódtak az á*dáz + háború* kollokációhoz, alacsony gyakorisággal:

- *ádáz háború kitör/ követel (áldozatot)*
- *ádáz háborút visel*
- *ádáz háborúra kelnek*

4.1.9 Az *ádáz* (lemma) + *düh* (lemma) kollokáció

Ebben a kollokációban a *düh* lemma nem szerepel összetett szó elemeként az *ádáz* környezetében.

Nem közvetlenül követik egymást a kollokáltak a következő kollokációkban: *ádáz, primitív düh; ádáz, testvéri düh*. A találati listában további három, nem azonos frázisban szereplő, tehát érvénytelen találat szerepelt.

5. függelék

Az *ádáz* további leggyakoribb kollokáltjai a Magyar Történeti Korpuszban

5.1.1 Az *ádáz* (lemma) + *harc* (lemma) kollokáció

Ebben a kollokációban a *harc* önálló lemmaként 6 esetben, összetett szó második tagjaként 2 esetben szerepel (mindkettőször *ádáz kenyérharc*).

A táblázatban szereplő igéken kívül az alábbi igék kapcsolódtak az á*dáz* + *harc* kollokációhoz:

- *ádáz harcot hirdet/lát/megsejt*
- *ádáz harcba beleveti magát*

5.1.2 Az *ádáz* (lemma) + *ellenfél* (lemma) kollokáció

Az *ellenfél* lemma nem szerepel összetett szó elemeként az *ádáz* környezetében. Mindhárom esetben közvetlenül követik egymást a kollokáltak.

5.1.3 Az *ádáz* (lemma) + *küzdelem* (lemma) kollokáció

Ebben a kollokációban a *küzdelem* lemma nem szerepel összetett szó elemeként az *ádáz* környezetében. Mindhárom esetben közvetlenül követik egymást a kollokáltak. A táblázatban szereplő igéken kívül az alábbi igék kapcsolódtak az á*dáz* + *küzdelem* kollokációhoz:

- *ádáz küzdelem indul*
- *ádáz küzdelmet színre visz*

5.1.4 Az *ádáz* (lemma) + *(kártya)csata* (lemma) kollokáció

Ebben a kollokációban a *csata* egyszer önálló lemmaként, egyszer a *kártyacsata* összetett szó második tagjaként szerepel. A táblázatban szereplő igén kívül az alábbi ige kapcsolódott az á*dáz* + *csata* kollokációhoz:

- *ádáz kártyacsatánál imbolyog*