SZIRMAI MONIKA

BEVEZETÉS A KORPUSZNYELVÉSZETBE

SEGÉDKÖNYVEK
A NYELVÉSZET TANULMÁNYOZÁSÁHOZ XLVI.

SZIRMAI MONIKA

BEVEZETÉS A KORPUSZNYELVÉSZETBE

A korpusznyelvészet alkalmazása az anyanyelv és az idegen nyelv tanulásában és tanításában

TINTA KÖNYVKIADÓ
BUDAPEST, 2005

SEGÉDKÖNYVEK
A NYELVÉSZET TANULMÁNYOZÁSÁHOZ XLVI.

Sorozatszerkesztő:
KISS GÁBOR

Lektor:
ANDOR JÓZSEF

ISSN 1419-6603
ISBN 963 7094 42 3

A kiadásért felel
a TINTA Könyvkiadó igazgatója
Felelős szerkesztő: Temesi Viola
Műszaki szerkesztő: Bagu László

TARTALOM

Ábrák jegyzéke 9

Táblázatok 11

Köszönetnyilvánítás 12

Kinek szól ez a könyv? 13

Bevezetés 15

1. Mi a korpusznyelvészet? 17

1.1. Bevezetés 17
1.2. Mi a korpusz? 17
1.3. A korpusz tervezése 23
1.3.1. A reprezentativitás 23
1.3.1.1. Mintavétel 23
1.3.1.2. A korpusz mérete 27
1.3.2. A korpuszok fajtái 32
1.3.2.1. A mintavétel módja szerint 32
1.3.2.2. A korpusz felhasználásának módja szerint 32
1.3.3. Jogi problémák 36
1.3.4. Átírás és annotáció 37
1.3.4.1. A beszéd átírása 37
1.3.4.2. A standard annotáció 38
1.3.4.3. Speciális annotációk 43
1.4. Összefoglalás 45

2. Számítástechnika és nyelvtudomány 47

2.1. Bevezetés 47
2.2. A számítástechnika fejlődése 47
2.3. A korpuszok fejlődése 50
2.3.1. A szellemi háttér 51
2.3.2. A korpusznyelvészet és a kapcsolódó tudományágak 53
2.3.2.1. A számítógépes nyelvészet 53
2.3.2.2. A mesterséges intelligencia 53
2.3.2.3. A számítógépes nyelvészet kutatási területe 54
2.3.3. A magyarországi számítógépes nyelvészetről 55
2.4. Folyóiratok 58
2.5. Összefoglalás 59

3. A korpuszokról ... 60
3.1. Bevezetés ... 60
3.2. Az elektronikus korpuszok előfutárai ... 60
3.2.1. A Szerb Nyelv Korpusza ... 60
3.2.2. A SEU Korpusz (Survey of English Usage Corpus) ... 61
3.3. A Brown Korpusz (1964) ... 63
3.4. A LOB Korpusz ... 63
3.5. A COBUILD projekt ... 64
3.6. A Brit Nemzeti Korpusz – British National Corpus (BNC) ... 66
3.7. Az Angol Nyelv Nemzetközi Korpusza (International Corpus of English – ICE) ... 67
3.8. A nem anyanyelvi angol korpuszok ... 69
3.8.1. A Longman Angol Nyelvtanulói Korpusz – Longman Corpus of Learners' English (LCLE) ... 70
3.8.2. A Nemzetközi Angol Nyelvtanulói Korpusz (International Corpus of Learners' English (ICLE)) ... 70
3.8.3. A Hongkongi Műszaki és Természettudományi Egyetem Angol Tanulói Korpusza (Hong Kong University of Science and Technology [HKUST] Corpus of Learner English) ... 72
3.8.4. Japán diákok angol nyelvű korpuszai ... 72
3.8.5. A Janus Pannonius Tudományegyetem Korpusza ... 72
3.8.6. Az Eötvös Loránd Tudományegyetem Korpusza ... 73
3.9. A korpuszok nyelvenként ... 73
3.9.1. További angol nyelvű korpuszok ... 74
3.9.1.1. A Brown Korpusz klónjai ... 74
3.9.1.2. Könyvkiadók korpuszai ... 75
3.9.1.3. Történeti nyelvészeti korpuszok ... 77
3.10. Az angol nyelvű korpuszok áttekintése ... 78
3.11. Magyar nyelvű korpuszok ... 80
3.11.1. A Magyar Nemzeti Szövegtár (MNSZ) ... 80
3.11.2. A Magyar Irodalmi és Köznyelv Nagyszótárának korpusza / Magyar Történeti Korpusz ... 81
3.11.3. Szeged Korpusz (http://www.inf.u-szeged.hu/projectdirs/hlt/) ... 86
3.11.4. Magyar dalszövegek ... 87
3.11.5. CHILDES Database: http://childes.psy.cmu.edu/ magyar nyelvű korpusza ... 87
3.11.6. A Hunglish Korpusz ... 87
3.11.7. A Magyar Webkorpusz ... 87
3.12. Egyéb nyelvek korpuszai ... 88
3.12.1. Német nyelvű korpuszok ... 89
3.12.1.1. A negr@ korpusz ... 89
3.12.1.2. A Tiger Korpusz ... 89
3.12.1.3. Freiburger Korpus ... 90
3.12.1.4. Dialogstrukturenkorpus ... 90
3.12.1.5. Pfeffer-Korpus ... 90

3.12.1.6. Telefonbeszélgetések (Brons-Albert, 1984) ... 90
3.12.2. Francia nyelvű korpuszok ... 91
3.12.2.1. PAROLE Francia Korpusz http://www.elda.org/catalogue/en/text/W0020.html ... 91
3.12.2.2. Francia Beszélt Nyelvi Korpusz ... 91
3.12.2.3. Kanadai Francia Korpusz ... 91
3.12.2.4. Le corpus VALIFLOUI (Variétés Linguistiques du Français en Louisiane) http://languages.louisiana.edu/French/Valifloui.html University of Louisiana at Lafayette ... 92
3.12.2.5. Le Corpus du Théâtre religieux français du Moyen Âge (Középkori Francia Vallásos Színház Korpusza) ... 93
3.12.3. A Szerb Nyelv Korpusza ... 93
3.12.4. A horvát nyelv korpusza ... 95
3.12.5. Szlovén nyelvű korpuszok ... 97
3.12.5.1. Szlovén – FIDA ... 97
3.12.5.2. BESEDA ... 97
3.12.6. Cseh nyelvű korpuszok ... 98
3.12.7. Lengyel korpuszok ... 98
3.13. Összefoglalás ... 98

4. A szoftverekről ... 100

4.1. Bevezetés ... 100
4.2. A korpuszok készítésekor használt programok ... 100
4.3. A konkordanciaprogramok ... 102
4.3.1. A kezdet kezdetén ... 110
4.3.2. Internetes felületen futó ingyenes programok ... 111
4.4. Konkordanciák készítése ... 112
4.4.1. Az MLCT ... 115
4.4.2. Simple Concordance Program SCP ... 120
4.4.3. ConcApp ... 126
4.4.4. AntConc ... 127
4.5. Összefoglalás ... 129

5. Korpusznyelvészeti módszerek az oktatásban ... 130

5.1. Bevezetés ... 130
5.2. Konferenciák és publikációk ... 130
5.3. Számítógéppel és nélküle ... 132
5.4. A késztermékek ... 133
5.4.1. Az egynyelvű tanulói szótárakról ... 133
5.4.1.1. Bevezetés ... 133
5.4.1.2. Anyanyelvi szótár – tanulói szótár ... 134
5.4.2. COBUILD kiadványok ... 139
5.4.2.1. Tankönyv ... 139
5.4.2.2. Segédanyagok ... 139

5.4.3. A Longman Grammar of Spoken and Written English (LGSWE) 142
5.4.4. Touchstone – új korpusz alapú tankönyv 143
5.5. Saját készítésű feladatok 144
5.5.1. Konkordanciák nyomtatásban 144
5.5.2. A „számok tükrében” 156
5.6. Számítógépes feladatok 157
5.7. Összefoglalás 162
A könyvben szereplő nyelvi korpuszok, szövegtárak és adatbázisok 164
Korpusznyelvészeti alapfogalmak 167
Bibliográfia 175
Tárgymutató 184
Névmutató 186
Korpuszok mutatója 189

ÁBRÁK JEGYZÉKE

Első fejezet
1. WordNet kererés eredménye 21–22
2. A Brown Korpusz összetétele 24
3. Az informatív próza alkategóriái a Brown Korpuszban 25
4. A szépprózа alkategóriái a Brown Korpuszban 25
5. A típusok és a hapaxok növekedése a COBUILD Korpuszban 29
6. Az IBM, APHB és a Hansard Korpusz lexikonjának növekedése (McEnery és Wilson, 1996: 157) 30
7. A Magyar Elektronikus Könyvtár figyelmeztető szövege 36–37
8. Prozódikus átirat 37
9. A BNC annotált szövege 39–40
10. A BNC szövege annotálás nélkül 40
11. Példa a csontváz/sekély elemzésre a UCREL honlapjáról http://www.comp.lancs.ac.uk/computing/research/ucrel/annotation.html 41
12. Példa az ágrajzra 42
13. Szemantikai annotáció (Leech 1997: 13) 43
14. Diskurzus annotáció (Leech 1997: 13) 43
15. Egy japán mondat morfológiai elemzése 44
16. Morfológiai elemzés vizuális elrendezésben 45

Második fejezet
17. Az ENIAC 2 48
18. A technológiával támogatott valós időben történő kommunikáció 55

Harmadik fejezet
19. Đorđe Kostić a korpusz anyagával a háttérben (1959) 61
20. Mondattani elemzés ágrajza az ICE-ben (http://www.ucl.ac.uk/english-usage/ice/annotate.htm) 68
21. Az MNSZ szövegtípusainak szövegszó szerinti aránya 80
22. Magyar Irodalmi és Köznyelv Nagyszótárának Korpusza / Magyar Történeti Korpusz kereső felülete 82
23. Találatok konkordancia formában a Magyar Irodalmi és Köznyelv Nagyszótárának Korpusza / Magyar Történeti Korpuszban 82–83
24. Bővebb kontextus és előfordulás helye 83
25. Előfordulás és rövid bibliográfia 84
26. Keresés eredménye teljes bibliográfiával 84–85
27. Ivo Andrić egy művének elemzése 95
28. Keresés eredménye a horvát korpuszban 96

Negyedik fejezet
29. *„Bornemissza”* konkordanciája az *Egri csillagok*ból [AntConc programm (Anthony, 2004]) 103
30. Konkordanciák mondat formában (*WordSmith* program) 104
31. Gyakoriság szerinti szólista az *Egri csillagok*ból (*WordSmith* programm) 105
32. Ábécé szerinti szólista (*WordSmith* program) 105

33. Statisztikai adatok az *Egri csillagok* 3 különböző fájljáról 107
34. A török* szövegkörnyezetében előforduló szavak (*WordSmith*) 108
35. *Web Concordancer* http://www.edict.com.hk/concordance/default.htm 112
36. Az *MLCT* program kezdő ablaka 115
37. *MLCT* fájl menüje 116
38. A Tools menü pontjai 117
39. 2n-gram számokkal szemléltetve 117
40. 3n-gram számokkal szemléltetve 118
41. A Collocation Parameters almenüje 118
42. Az Extract Collocates By statisztikai együtthatókat felkínáló alpontjai 118
43. Kontroll ikonok és szövegablakok 119
44. A programablak konkordanciákkal a jobb oldalon 119
45. Új projekt létrehozása 121
46. A karakterkészlet szerkesztőablaka 122
47. A projekt kezdő ablaka 123
48. Statisztikai adatok 124
49. A pojektstatisztika adatai 125
50. Kulcsszavak szerkesztése 125
51. Teszt funkció a *ConcApp* programban 126
52. Az *AntConc* kezelőfelülete 127
53. A keresett szó elhelyezkedése a szövegben 128
54. Szócsoport-keresés 128

Ötödik fejezet

55. A *beaver* címszó az anyanyelvi szótárban 136
56. A *beaver*[1] címszó az anyanyelvi gyermekszótárban 137
57. A *beaver* meghatározása a *Longman Dictionary of Contemporary English* szerint 137
58. A *beaver* meghatározása az *Oxford Advanced Learner's Dictionary* szerint 137
59. A *beaver* meghatározása a *CollinsCOBULD* szerint 138
60. John Sinclair az ICAME 25 Konferencián Veronában 141
61. Átlagos szószám szerkezetenként a különböző regiszterekben 143
62. A *vár* konkordanciája az MNSZ-ből 145
63. Gyakorlat az egyeztetésre (Tim Johns) 145–146
64. Gyakorlat a névelők használatára (Tim Johns) 146
65. A *vörös* keresési eredménye konkordanciapéldákkal az MNSZ-ben 148–149
66. A *piros* eredménye konkordanciapéldákkal az MNSZ-ben 150–151
67. Konkordanciák a keresett szó nélkül 151–152
68. A *flat* konkordanciája 152–153
69. Feladat a szuperordinátok és hiponímák gyakorlására 153
70. A *bánat* és *szerez* lekérdezése az MNSZ-ből 154
71. A *bánat* és *okoz* lekérdezése az MNSZ-ből 154–155
72. A verursach szemantikai prozódiája (Bill Dodd honlapjáról) 155
73. A fordítási ekvivalensek egyik ábrázolási módja 156
74. A *vörös* és a *veres* gyakoriságának összehasonlítása 156–157
75. Konkordanciák szerkesztése a Contexts programhoz 159
76. Utasítások szerkesztése a Contexts programhoz 159
77. A Contexts.ini fájl tartalma 160
78. Francia–angol párhuzamos szöveg 161
79. A *tous* konkordanciái és fordításai 161
80. Parallel szövegek gyakorisági listája 162

TÁBLÁZATOK

Első fejezet
1. Relációs adatbázis ... 20
2. Az ICE összetétele a http://www.ucl.ac.uk/english-usage/ice/design.htm# honlap alapján ... 26
3. A korpusz mérete és jellemzői Krishnamurthy (1997b) alapján ... 28
4. A Magyar Nemzeti Szövegtár alapkódjai ... 39
5. A Szerb Nyelv Korpuszának elemzett formája ... 41

Második fejezet
6. A kereskedelmi forgalomban levő elektronikus számítógépek első 30 éve: technikai összehasonlítások ... 49
7. A korpusz és az elméleti nyelvész paródiája (Lager 1995: 3) ... 51–52

Harmadik fejezet
8. A SEU írott eredetű szövegei ... 62
9. A SEU beszélt nyelvi szövegei ... 62
10. A Bank of English szerkezete Krishnamurthy alapján (1997a: 79) ... 65
11. Az ICE brit alkorpuszának összetétele (http://www.ucl.ac.uk/english-usage/ice-gb/sampler/download.htm alapján) ... 69
12. Az ICLE alkorpuszai (http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm alapján) ... 71
13. Példa a FLOB Korpuszban szereplő szövegek adataira (http://helmer.aksis.uib.no/icame/flob/kata.htm alapján) ... 74–75
14. Példa a FROWN korpuszban szereplő szövegek adataira (http://khnt.hit.uib.no/icame/manuals/frown/KATA.HTM alapján) ... 75
15. A Cambridge Nemzetközi Korpusz http://uk.cambridge.org/elt/corpus/cic.htm alapján ... 76
16. Angol nyelvű korpuszok összefoglaló táblázata ... 78–80
17. A Szeged Korpusz összefoglaló adatai ... 86
18. A Hunglish Korpusz mérete ... 87
19. A Webkorpusz mérete (http://lab.mokk.bme.hu/eszkozok/webkorpusz/) ... 88
20. A VALIFLOUI Korpusz adatai ... 92–93
21. Példa a tartalom felsorolására ... 94
22. A XII-től a XVIII. századig terjedő szövegek korpuszának adatai a szavak számával együtt ... 94–95
23. A Cseh Nemzeti Korpusz ... 98

Negyedik fejezet
24. Ingyenesen letölthető programok ... 113

Ötödik fejezet
25. A korpuszok oktatásban való használatával foglalkozó konferenciák ... 131

KÖSZÖNETNYILVÁNÍTÁS

E könyv soha nem jöhetett volna létre Cseresnyési László biztatása és segítsége nélkül. Szintén neki köszönhetem Kiss Gáborral, a Tinta Könyvkiadó igazgatójával való megismerkedésem, melynek eredményeként könyvem itt került kiadásra.

Egy könyv megjelenése nem csak a szerző érdeme, még akkor sem, ha a címlapon csak az ő neve szerepel. Mivel ez az első magyar nyelven megjelenő írásom, úgy érzem, hogy ez a legjobb alkalom arra, hogy nyilvánosan köszönjem meg a segítséget mindazoknak, akik nélkül nem lettem volna azzá, aki vagyok. Középiskolás koromtól kezdve sok kiváló szakember irányított és segített nyelvi és nyelvészeti ismereteim megszerzésében, így lehetetlen lenne mindenkit felsorolni, annak ellenére, hogy az én emlékezetemben mind élénken élnek. Vannak azonban olyanok is, akik az életem alakulására is nagy hatást gyakoroltak.

Pályaválasztásomban döntő szerepet játszott két középiskolai tanárom: Albert Sándor és Novák György. Most is hálás vagyok akkori tanácsaikért. Debreceni egyetemi éveim alatt az angol és a francia tanszék minden tanára nagy hatással volt szellemi fejlődésem alakulására. A Birminghami Egyetemen a British Council ösztöndíjasaként eltöltött egy év alatt pedig olyan kutatókkal kerültem napi kapcsolatba, mint John Sinclair, Susan Hunston, Ramesh Krishnamurthy, Dave és Jane Willis. Szintén hálával tartozom Hollósy Bélának, aki doktori tanulmányaim alatt elfoglaltsága ellenére is mindenben és mindenkor segítségemre sietett. Doktori tanulmányaim alatt ismerkedtem meg Andor Józseffel, aki hihetetlen energiával és lelkesedéssel végzi mind pedagógusi, mind kutatói munkáját. E könyvet nem csak kész állapotában lektorálta, hanem azon túlmenően, az írás megkezdésétől kezdve készségesen válaszolt kérdéseimre, és segített tanácsaival. Fenyvesi Anna és Heitzmann Judit a kéziratra vonatkozó észrevételeikkel segítették munkámat.

A szakmai segítség mellett azonban elengedhetetlen a család és a barátok türelme, támogatása és biztatása. Köszönöm szüleimnek, hogy mindig mindenben mellettem álltak és segítettek. A bajai Kerényi és Rácz családnak, a hiroshimai Szaszaki Tadaszunak és Murakami Midorinak külön köszönettel is tartozom. Számítógépes problémáim megoldásában Rácz Györgyre, Schmíz Istvánra, Oláh Gyulára és Varjú Tiborra bármikor számíthattam.

Ha e sok segítség és támogatás ellenére mégis előfordulnak hiányosságok vagy pontatlanságok e könyvben, akkor az csakis rajtam múlott.

Hiroshima, 2005 tavasza

Szirmai Monika

KINEK SZÓL EZ A KÖNYV?

Könyvemnek az a célja, hogy a lehető legszélesebb közönséggel ismertesse meg a korpusznyelvészet alapjait és alkalmazási lehetőségeit. A könyv írásakor igyekeztem egyaránt szem előtt tartani a nyelvszakos egyetemi hallgatókat, az általános és középiskolai tanárokat, a magyar nyelvet és irodalmat vagy idegen nyelvet tanító nyelvtanárokat, valamint – kortól függetlenül – a nyelvtanulókat. Arra törekedtem, hogy a nyelvészeti zsargont elkerülve vagy megmagyarázva, egyszerű, könnyen olvasható formába öntsem gondolataimat.

A korpusznyelvészet számos területen alkalmazható, de számomra elsősorban az a tény fontos, hogy jelentősen elősegítheti mind az anyanyelvvel kapcsolatos ismereteink bővítését és pontosítását, mind az idegen nyelvek elsajátítását. A nyelvtanároknak vagy az anyanyelvi beszélőknek segítséget nyújthat olyan kérdések megválaszolásában, amelyekre a diákok a nyelvtankönyvekben nem találnak választ. Munkám megjelentetését az is indokolja, hogy magyar nyelvű könyv még nem készült a korpusznyelvészetről, így csak idegen nyelven, elsősorban angolul olvasható ezzel foglalkozó szakirodalom.

BEVEZETÉS

Néha az ember életét egy-egy találkozás örökre megváltoztathatja. Ilyen fordulópont volt az én életemben az 1993–94-es tanév, amikor felnőtt fejjel ismét diákként ültem a padban, és egy előadást hallgattam, vagyis inkább egy számítógépes program bemutatóját néztem. A program neve Contexts volt (ebben a könyvben is bemutatásra kerül), az előadó pedig a Birminghami Egyetem tanára, Tim Johns. Hogy miért éppen egy számítógépes program hatott rám ilyen elemi erővel, amikor számos elméleti előadáson már túl voltam minden különösebb lelkesedés nélkül? Bizonyára az előadó személyes varázsa is közrejátszott, de elsősorban a játék izgalma és a szellemi kihívás ragadott meg mint nyelvtanulót. Mint nyelvtanárnak, azonnal az alkalmazási lehetőségek sora futott végig a fejemben, hiszen ha nekem ennyire tetszett, talán a diákjaimnak is hasznos és szórakoztató lehet.

A nagy lelkesedés nem hozott olyan gyors előrelépést, mint azt gondolhatnánk, mert sajnos abban az időben sem számítógéppel, sem pedig informatikai ismeretekkel nem rendelkeztem. Szeretném megnyugtatni az olvasót, hogy ma már nincs is szükség olyan programozási ismeretekre, mint a Windows 95 operációs rendszer megjelenése előtt, és a számítógép kezelése is egyre könnyebb lett. E könyv írásakor azonban feltételezem, hogy az olvasó rendelkezik alapvető számítógépes ismeretekkel, azaz tudja, hogy hogyan kell megnyitni egy könyvtárt vagy fájlt, tud menteni, el tud indítani egy programot, tudja kezelni a menürendszert stb.

A modern korpusznyelvészet feltételezi a számítógép használatát, de nem azonos a számítógépes nyelvészettel annak ellenére, hogy számos átfedés található a kettő között. A kutatás céljai és az eredmények felhasználásának lehetőségei is különböznek. A programozási ismeretek a korpusznyelvészet művelésekor is sokat segíthetnek, de a számítógépes nyelvészetben elengedhetetlenek.

A nyelvtanulás számomra olyan, mint a felfedezés. Akár írott szövegről, akár élőbeszédről legyen szó, ha magam „fedezem fel” a szabályt vagy veszek észre valamit, az sikerélményt nyújt, és sokáig megmarad az emlékezetemben. A mások által közölt vagy könyvben olvasott szabályokra azonban már nem emlékszem olyan könnyen és hamar el is felejtem őket. A diákokat is kutatóknak tekintem. Minél magasabb szinten szeretnének elsajátítani egy nyelvet, annál jobb és önállóbb kutatókká kell válniuk. A diákoknak három nagyon fontos dologra van szükségük: egyrészt megfigyelőképességre, másrészt képesnek kell lenniük helyes következtetések levonására a rendelkezésre álló adatok alapján, harmadrészt pedig képesnek kell lenniük intelligens találgatásra. Egy egyszerű példával szeretném ezt bemutatni.

Tegyük fel, hogy valaki hirtelen olyan környezetbe kerül, amelynek nyelvét (mondjuk a japánt) nem beszéli. Mivel élénken figyel, egy idő után észreveszi, hogy minden étkezés előtt hallja az *itadakimasz* kifejezést, étkezés után pedig azt, hogy *gocsiszó-*

szamadesta. Ennek alapján levonhatja a következtetést, hogy ezt a két kifejezést akkor használják, amikor a magyarban a *Jó étvágyat!* és az *Egészségünkre!* fordulatokat. Az *itadakimasz* kifejezés jelentését a helyzetből fakadóan a *Jó étvágyat!* kifejezéssel fogja azonosítani mindaddig, amíg új, más helyzetben nem találkozik ugyanezzel a kifejezéssel, ahol nyilvánvalóan más jelentésben használják.

Azok a nyelvtanulók lesznek sikeresek, akik az ilyen helyzeti megfigyelőképességüket rendszeresen használják. Sokan automatikusan teszik ezt, másoknak fel kell erre hívni a figyelmét, és még gyakorlásra szoruló tanulók is vannak. De ezek a képességek gyakorlással fejleszthetők. Tapasztalataim szerint erre különösen alkalmasak a korpusznyelvészet szellemében készült feladatok.

Ez a könyv tehát az olvasók széles skálájának igényeit igyekszik kielégíteni – a szakembertől kezdve az érdeklődő diákig. Öt fejezetből áll. Az első a korpusznyelvészet alapjait mutatja be. Mivel a számítástechnika fejlődése meghatározó volt a korpusznyelvészet szempontjából, a második fejezetben a számítástechnika és nyelvészet kapcsolatáról esik majd szó. Ezt követi a legfontosabb korpuszok bemutatása a harmadik fejezetben. A negyedik fejezet a korpuszok használatához szükséges számítógépes programokat ismerteti. A legtöbb figyelmet e fejezetben az úgynevezett konkordanciaprogramoknak szenteljük: illusztrációk és részletes leírás segítségével szeretném lehetővé tenni, hogy még az angolul nem tudó érdeklődők is kezelhessék e programokat. Az ötödik fejezet a korpusznyelvészet oktatásban való felhasználásának lehetőségeit mutatja be. Számos mintafeladatot és alkalmazási ötletet tartalmaz, melyeket bármely nyelv tanításakor vagy tanulásakor fel lehet használni. Eddigi kutatásaim és tanítási tapasztalataim, valamint a korpusznyelvészeti szakirodalom jelentős többsége az angol nyelvre vonatkozik, így annak ellenére, hogy igyekeztem minél több esetben magyar nyelvű példát is hozni, azok nagy része angol nyelvű. Remélem, hogy így is haszonnal forgatja majd e könyvet minden olvasója.

1. MI A KORPUSZNYELVÉSZET?

1.1. Bevezetés

„Nem túlzás, ha azt állítjuk, hogy az utóbbi néhány évtizedben a korpuszok és a korpuszok tanulmányozása forradalmasította a nyelv és a nyelv alkalmazásának tanulmányozását."[1], írja könyve bevezetőjében Susan Hunston (2002). Ennek ellenére még néhány évvel ezelőtt is megtörtént, hogy nyelvtanárok és nyelvészek zavartan és értetlenül néztek, amikor a *korpusznyelvészet* kifejezést hallották. Sokan közülük még soha nem hallották ezt a kifejezést, néhányuknak derengett valami, de egyikük sem tudta, hogy pontosan mi is az. A számítógép elterjedésével, és az internethez való hozzáférési lehetőség rohamos javulásával azonban a legtöbb szakember ma már ismeri a *korpusz* és a *korpusznyelvészet* kifejezést. Manapság a nyelvtanulóknak szóló tankönyvek, szótárak és egyéb kiadványok egyik elvárása az, hogy korpuszra és korpuszelemzésekre épüljenek.

Ebben a fejezetben először a korpusz meghatározására valamint a korpusztervezés és készítés problémáinak (reprezentativitás, a mintavétel, méret, célkitűzések és jogi problémák) bemutatására kerül sor. Ezt követi a korpuszok annotációjának[2] tárgyalása, és egy általános áttekintés zárja a fejezetet.

1.2. Mi a korpusz?

Maga a *korpusznyelvészet* kifejezés akkor vált szakkifejezéssé, amikor ezzel a címmel jelent meg egy tanulmánygyűjtemény Jan Aarts és Willem Meijs szerkesztésében 1984-ben. Ennek ellenére a szakszótárakban még az 1990-es évek végén sem szerepelt a korpusznyelvészet szó, bár majdnem mindegyikben fellelhető volt a *korpusz* szó meghatározása.

Magyar nyelven nem jelent még meg komolyabb nyelvészeti szakszótár, így csak a *Nyelvi fogalmak kisszótárában* találhattam meg a meghatározását, mely szerint a *korpusz* „meghatározott szempontok alapján kiválasztott szövegmennyiség, amelyen a nyelvész vizsgálatát végzi" (Kugler & Tolcsvai Nagy, 2000: 132). E meghatározást alapul véve akár egy mondat vagy egy szó is korpusznak tekinthető, pedig ennél sokkal nagyobb mennyiséget tekint korpusznak a szakirodalom. A másik komoly probléma ezzel a meghatározással az, hogy csak a mennyiségről beszél, és említést sem tesz a tartalomról, a tárolás módjáról vagy egyéb jellemzőkről.

---

[1] Eredetiben: *"It is no exaggeration to say that corpora, and the study of corpora, have revolutionised the study of language, and of the applications of language, over the last few decades."*

[2] A szövegfeldolgozás során a szövegbe illesztett, és a szövegre vonatkozó információt nevezzük annotációnak (lásd 1.3.4. rész).

A hazai korpuszkutatás központjának, a Magyar Tudományos Akadémia Nyelvtudományi Intézete Korpusznyelvészeti Osztályának honlapján a következő meghatározást találjuk (http://corpus.nytud.hu/mnsz/bevezeto_hun.html):

> A *korpusz* ténylegesen előforduló írott, vagy lejegyzett beszélt nyelvi adatok gyűjteménye. A szövegeket valamilyen szempont szerint válogatják és rendezik. Nem feltétlenül egész szövegeket tartalmaz, és nem csak tárháza a szövegeknek, hanem tartalmazza azok bibliográfiai adatait, bejelöli a szerkezeti egységeket (bekezdés, mondat). Emellett pedig feltünteti a szavak mellett szófaji kódjukat is.

Ez a meghatározás igen pontos és alapos. Minden fontos ismérv szerepel benne, kivéve talán kettőt, amelyet esetleg evidenciaként kezeltek, így azokat meg sem említi. Az egyik az, hogy a modern korpuszokat elektronikus formában tárolják, a másik pedig az, hogy a nyelv vizsgálatának céljából hozzák létre. Mint a következőkben látni fogjuk, a korpusz szónak számos jelentése van, így sokféleképpen lehet értelmezni. Szinte minden korpusznyelvészettel foglalkozó műben szerepel valamilyen meghatározás. Nézzünk meg néhányat az angol szakirodalomból, hiszen ez a legbővebb. Tom McArthur (1992: 265–266) így határozza meg a korpuszt szócikkében:

> KORPUSZ [13. század: latinból *corpus*, test. A többes száma általában *corpora*]. (1) szövegek gyűjteménye, különösen, ha teljes és önálló egység: *az angolszász versek korpusza*. (2) Többes száma *corpuses* is lehet. A nyelvészetben és lexikográfiában az általában elektronikus adatbázisként tárolt, egy adott nyelvre többé-kevésbé reprezentatívnak tekinthető írott szövegek, szóbeli közlések vagy egyéb minták gyűjteménye. Jelenleg a számítógépes korpusz több millió szót tárolhat, amelyek tulajdonságait *címkézéssel* (azaz a szavakat és más kifejezéseket azonosító és osztályozó címkével látják el), valamint konkordancia programok segítségével elemezhetik. A *korpusznyelvészet* az adatok ilyen korpuszban való tanulmányozását végzi.[3]

Mielőtt tovább mennénk, néhány megjegyzés elkerülhetetlen a fenti idézettel kapcsolatban. Először is, mind az (1) és a (2) jelentésben a többes számok az angol nyelvhasználatra vonatkoznak. Az angol szakirodalomban a *collection of texts* (*szövegek gyűjteménye*) és a *body of texts* (szövegek halmaza, tára) kifejezés is gyakran szerepel. A magyar fordításban mindkét esetben *gyűjtemény*ként szerepel, jóllehet a *body of texts* hallatán az egyet alkotás, az összetartozás érzése talán erősebb, mint a *collection of texts* esetében. Erre jó példa az is, hogy a *body* magyar fordítása különböző kifejezésekben sokszor a *szervezet*, *testület, csoport* szóval történik.

[3] Eredetiben: *"CORPUS [13c: from Latin corpus body. The plural is usually corpora]. (1) A collection of texts, especially if complete and self-contained: the corpus of Anglo-Saxon verse. (2) Plural also corpuses. In linguistics and lexicography, a body of texts, utterances, or other specimens considered more or less representative of a language, and usually stored as an electronic database. Currently, computer corpora may store many millions of running words, whose features can be analysed by means of tagging (the addition of identifying and classifying tags to words and other formations) and the use of concordancing programs. Corpus linguistics studies data in any such corpus."*

David Crystal (1992: 410) szerint a korpusz *„a nyelv reprezentatív mintája, amelyet nyelvészeti elemzés céljából állítottak össze”*.[4] Nelson Francis (1982: 7), a modern korpusznyelvészet egyik úttörője a nyelvi korpuszt *„az adott nyelvre, dialektusra vagy más nyelvi alcsoportra nézve reprezentatívnak tekintett szövegek gyűjteménye”*[5]-ként határozza meg. John Sinclair (2001) pedig úgy mutat rá a szövegek gyűjteménye és a korpusz közötti különbégre, hogy kiemeli azt, hogy a korpuszt a nyelvészek „nyelvészeti státusszal” ruházzák fel, azaz nyelvészeti vizsgálatok végzésére alkalmasnak tekintik, hiszen a szövegek gyűjtése nem esetleges módon, hanem bizonyos külső kritériumok alapján történik. Sinclair itt utal Clear (1992) cikkére, mely ezeket a kritériumokat taglalja, majd hozzáteszi, hogy a korpusz implicit módon magában hordozza azt a feltételezést, hogy a *„benne használt nyelv belső mintáinak vizsgálata nyelvészeti szempontból tanulságos és eredményes lesz”*[6] (J. M. Sinclair, 2001: xi).

Ezek a meghatározások többé-kevésbé hasonlítanak egymáshoz. A bennük szereplő kulcsszavak: 1. *reprezentatív* (szövegek gyűjteménye), mely az utolsó idézetet kivéve mindegyik meghatározásban központi helyet foglal el, és amelyről a későbbiekben még szólni fogunk; 2. *elektromos formában tárolt* – ami az adatok kezelése miatt fontos; 3. elemzésük eredménye *nyelvészeti szempontból hasznos* lesz. Az 1. és a 3. pontból következik, hogy a korpusz létrehozása különös gondot igényel. Tehát, a gondosan összeválogatott szövegfájlok (a számítógépen fájlnév.txt) tára korpuszt képez. Nyelvtanárok vagy irodalomtanárok esetében például egy osztály vagy évfolyam egy adott témáról készített összes dolgozata, vagy egy teljes tanévben írt munkájának a gyűjteménye. A diák szemszögéből nézve az általa valaha is idegen nyelven írt dolgozatok gyűjteménye is korpuszt képezhet.

Miután meghatároztuk, hogy mi is a korpusz, nézzük meg, hogy mit nem tekinthetünk annak. Jóllehet az elektronikus szöveggyűjtemények és adatbázisok is szövegeket tartalmaznak, sőt még nyelvészeti elemzést is végezhetünk velük, mégsem nevezhetjük korpusznak ezeket, még akkor sem, ha a nevükben szerepel a korpusz kifejezés. Például az Oxfordi Szövegarchívum (Oxford Text Archive http://ota.ahds.ac.uk/) válogatás nélküli szövegek tára, mely 25 nyelven írt, több mint 2500 forrással rendelkezik, és folyamatosan gyarapodik. Magyar nyelvű szöveget nem tartalmaz, de például az alapvető 1945 japán írásjelet innen le lehet tölteni. Bárki, aki kedvet érez hozzá, elektronikus „adományával” hozzájárulhat az archívum fejlesztéséhez.

A Pennsylvaniai Egyetem ad otthont a Nyelvészeti Adatok Konzorciumának (Linguistic Data Consortium, LDC, http://www.ldc.upenn.edu/), melyet 1992-ben alapítottak azzal a céllal, hogy nyelvi adatbázisokat, szöveggyűjteményeket, nyelvészeti elemző programokat hozzon létre, gyűjtsön össze más forrásokból, és osszon meg a tagokkal és kutatókkal. A jelszavuk az, hogy „minél több, annál jobb”.[7] Ebből is kiderül, hogy semmilyen külső szempontok nem érvényesülnek az adatgyűjtésben. Természetesen az

[4] Eredetiben: *“representative sample of language, compiled for the purpose of linguistic analysis”*.
[5] Eredetiben: *“a collection of texts assumed to be representative of a given language, dialect, or other subset of a language”*.
[6] Eredetiben: *“an investigation into the internal patterns of the language used will be fruitful and linguistically illuminating”*.
[7] Eredetiben: “No data like more data.”

itt felhalmozott gyűjteményből gondos válogatással saját korpuszt hozhatunk létre, így az ilyen szöveggyűjtemények is nagyon hasznosak lehetnek a korpuszt használó vagy elemző diákok, tanárok és nyelvészek számára.

Az adatbázis kifejezést lehet tágabb vagy szűkebb értelemben venni. Tágabb értelemben a korpusz is adatbázis, hiszen az adatbázis olyan módon rendszerezett adatok gyűjteménye, amely lehetővé teszi az adatok gyors elérését, kezelését és megújítását. Ennek ellenére, jóllehet az adatbázisoknak több fajtája létezik, az adatbázis szó esetében szinte mindenki kivétel nélkül relációs adatbázisra[8] gondol, hiszen mindennapi életünkben egyre ritkábban találkozunk más fajtával. A relációs adatbázis lényege az, hogy az adatok közötti kapcsolatok nincsenek előre meghatározva, eltérően a hierarchikus vagy hálós adatbázisoktól. A relációs adatbázis hallatán gondoljunk egy táblázatra, ahol az összetartozó adatok a táblázat egy-egy sorában szerepelnek, az oszlopok nevei pedig a bennük található adat jellegére utalnak. A táblázat két sora soha nem lehet teljesen azonos. Közérthető módon ad bővebb felvilágosítást az adatbázisokról Nagy Attila írása az alábbi honlapon: http://www.kfki.hu/chemonet/hun/eloado/adatb/index.html, illetve dr. Siki Zoltán (1995) ismertetése a http://www.agt.bme.hu/szakm/adatb/adatb.htm lapon.

Nézzünk meg azért egy egyszerű példát (1. táblázat).

Sorszám	Szerző(k)	Cím	Kiadás éve	Kiadás helye	Kiadó	ISBN	Ár

1. táblázat: Relációs adatbázis

Ebben a táblázatban egy házikönyvtár adatait lehetne például tárolni, amelyben az egyes könyvekre vonatkozó összes információ egy sorban szerepel. Természetesen egy író több könyvet is kiadhat egy évben, azonos kiadónál, azonos helyen, és még az ára is lehet azonos. A sorszám, a cím és az ISBN szám, amely a könyvet egyértelműen azonosítja, azonban különböznek. Az adatbázist egyszerűen lehet bővíteni, például a vásárlás vagy tárolás helyével, vagy egyéb információval. Az egyes adatok és az oszlopok sorrendje tetszőleges és könnyen megváltoztatható.

A szövegek általában lineárisak, azaz az elejétől a végéig („sorrendben”) olvassuk őket. Mivel a korpusz is szövegeket tartalmaz, így a korpuszt is olvashatnánk lineárisan. Tulajdonképpen a számítógép segítségével ezt is tesszük, hiszen a gép nagy sebességgel végigpásztázza a szöveget, és a keresett szó vagy kifejezés összes előfordulását listaként mutatja a képernyőn. A szótárak és az adatbázisok azonban nem „lineáris olvasásra” készültek, hanem a szükséges információ gyors megkeresését szolgálják. Ha egy hétköznapi számítógépes példával szeretnénk illusztrálni a kétfajta működési elvet, akkor

[8] Az IBM-nél E. F. Codd (A Relational Model of Data for Large Shared Data Banks) találta fel a relációs adatbázist 1970-ben. A cikk első megjelenési helye: *Communications of the ACM*, Vol. 13, No. 6, June 1970, pp. 377–387. Reprint változatát azonban a http://www.acm.org/classics/nov95/toc.html címen is elérhetjük.

az MS Word vagy Word Perfect szövegszerkesztők jó példák lehetnek a lineáris működésre, az MS Access adatbázis program pedig a nem lineárisra. Aki készített már adatbázist, az minden bizonnyal nehezebb feladatnak tartotta, mint az egyszerű szövegszerkesztést. Ezek alapján könnyen beláthatjuk, hogy a korpuszok és adatbázisok között különbséget kell tennünk.

Az amerikai Princeton Egyetemen kifejlesztett WordNet http://www.cogsci.princeton.edu/~wn/ és európai mása, az Euronet on-line lexikális adatbázisok, amelyek már bizonyos nyelvi elemzések eredményeit felhasználva készültek, és részben korpuszok elemzésére is támaszkodnak. A WordNetről írt kb. 300 tételt tartalmazó bibliográfia a http://engr.smu.edu/~rada/wnb/ címen található meg az interneten, melyek közül a főbb művek a következők: Fellbaum (1998) és Miller et al. (1990). Mivel ezek az adatbázisok nem tekinthetők igazi korpusznak, így ezekről sem fogunk részletesen szólni, de a könnyebb érthetőség kedvéért álljon itt egy példa. A WordNet keresőjébe beírt angol *mind*[9] szó eredményeként a következőket kaptuk:

Overview for

"mind"

The **noun** "mind" has 7 senses in WordNet.

1. **mind**, head, brain, psyche, nous -- (that which is responsible for one's thoughts and feelings; the seat of the faculty of reason; "his mind wandered"; "I couldn't get his words out of my head")
2. **mind** -- (recall or remembrance; "it came to mind")
3. judgment, judgement, **mind** -- (an opinion formed by judging something; "he was reluctant to make his judgment known"; "she changed her mind")
4. thinker, creative thinker, **mind** -- (an important intellectual; "the great minds of the 17th century")
5. **mind** -- (attention; "don't pay him any mind")
6. **mind**, idea -- (your intention; what you intend to do; "he had in mind to see his old teacher"; "the idea of the game is to capture all the pieces")
7. **mind**, intellect -- (knowledge and intellectual ability; "he reads to improve his mind"; "he has a keen intellect")

Search for Synonyms, ordered by estimated frequency of senses

☑ Show glosses

☐ Show contextual help

Search

The **verb** "mind" has 6 senses in WordNet.

[9] Mivel a nyelvtanulók sokszor találkozhatnak egy idegen nyelv tanulása során olyan szavakkal, amelyek írásképe vagy kiejtése teljesen azonos egy magyar szóéval, nem tartottuk fontosnak, hogy a véletlen egybeesés miatt más példát válasszunk.

1. **mind** -- (be offended or bothered by; take offense with, be bothered by; "I don't mind your behavior")
2. **mind** -- (be concerned with or about something or somebody)
3. take care, **mind** -- (be in charge of or deal with; "She takes care of all the necessary arrangements")
4. heed, **mind**, listen -- (pay close attention to; give heed to; "Heed the advice of the old men")
5. beware, **mind** -- (be on one's guard; be cautious or wary about; be alert to; "Beware of telephone salesmen")
6. **mind**, bear in mind -- (keep in mind)

1. ábra: WordNet keresés eredménye

Az angolul tudó olvasók már első pillantásra láthatják, hogy nem „lineárisan olvasható” szöveget vagy szövegtöredékeket kaptunk eredményként, hanem egy felsorolást, mely a *mind* szó főnévi és igei használatban előforduló jelentéseit és szinonimáit összegzi. A WordNet adatbázisában a szavakat lexikális alapon szinonima halmazokba csoportosították és a kereséskor ezen kapcsolatok eredményeit kapjuk vissza. Az érdeklődők Unix és Windows változatban szabadon le is tölthetik az adatbázist, jelenleg a második változatot.

Érdemes itt megemlíteni, hogy Európában és az Egyesült Államokban egészen a közelmúltig igen különböző céllal hoztak létre korpuszokat. Európában többnyire nyelvészeti elemzések céljából készültek korpuszok, míg az óceán túlpartján inkább a technikai fejlődés elősegítése, például beszédfelismerés volt az elsődleges cél. A technikai jellegű kísérletekhez egyre több adatra, nem pedig külső szempontok alapján kiválogatott szövegek gyűjteményére volt szükség. John Sinclair megjegyzi, hogy *„az Egyesült Államok a Brown Korpusszal megkezdett kiváló indulás után húsz évig lemaradásban volt Európával szemben (mi több, az Amerikai Nemzeti Korpusz, amely a tíz évvel ezelőtti Brit Nemzeti Korpusz klónja, írásom idején még csak a tervezés szakaszában van.”*[10] (J. M. Sinclair, 2001:ix). Többek között ez az oka annak, hogy ebben a könyvben nagyrészt az Európában készült korpuszokat mutatom be.

Érdemes itt felhívni a figyelmet arra, hogy a korpuszok neve is adhat némi felvilágosítást magáról a korpuszról. Az előző bekezdésben szereplő Brown Korpusz a nevét arról az amerikai egyetemről kapta, ahol a korpuszt létrehozták. Sok esetben a korpuszt a projekben együttműködő intézményeknek otthont adó városokról nevezik el, például a Lancaster-Oslo/Bergen Korpusz. A város neve gyakran egyébként is része az intézmény hivatalos nevének. Gyakran utalnak még a korpusz jellegére is a nevében, például a *nemzeti* jelző arra utal, hogy a korpusz az adott nyelv jelenbeli állapota lehető legátfogóbb elemzésének készítéséhez kíván segítséget nyújtani, ezért a korpusznak a kortárs szövegek széles skáláját kell tartalmaznia. Így tehát sok esetben már a korpusz nevének hallatán következtethetünk a tartalmára is. A pontos névadás azonban azt eredményez-

[10] Eredetiben: *“The United States, after the brilliant start given by the Brown Corpus, then lagged behind Europe for twenty years (indeed the American National Corpus, cloned from the British one of ten years ago, is at the planning stage as I write).”*

heti, hogy a korpusz neve meglehetősen hosszú lesz, ezért gyakran csak az ezt rövidítő betűszót (akronimát) használják, pl. LOB, CANCODE, amint ezt az 1.3.2. részben, a korpuszok fajtáinak ismertetésekor látni fogjuk.

1.3. A korpusz tervezése

1.3.1. A reprezentativitás

A korpusz nem pusztán szövegek véletlen halmaza, hanem tudatosan megtervezett gyűjtemény, amelynek összeállításakor az ezen elvégezni kívánt nyelvi elemzést tartjuk szem előtt. Könnyen belátható, hogy az elemzésünk tárgyának, ez esetben a korpusznak, alkalmasnak kell lennie a kitűzött elemzés elvégzésére. Ha például az 1960-as évek nyelvét szeretnénk összehasonlítani az 1990-es évekével, két korpuszt kell összeállítanunk, amelyekben az egyik az 1960-as évek szövegeit, a másik pedig az 1990-es évek szövegeit tartalmazza. Ez a kutatási terv hatalmas mennyiségű adatot igényelne, hiszen a nyelv magában foglal mindent ebben a megfogalmazásban: a diákszlengtől kezdve a filozófiai értekezéseken át a mikrohullámú sütő használati utasításáig. Ha azt szeretnénk, hogy a korpusz valóban reprezentatív legyen, el kell döntenünk, hogy mit, illetve miből mennyit veszünk bele a korpuszba.

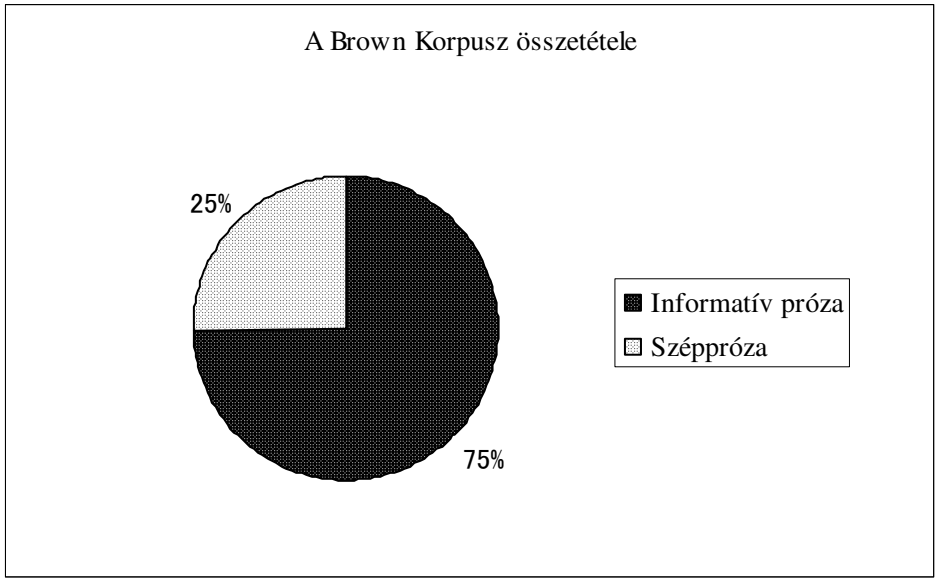
A legtöbb nyelvész eddig inkább az intuíciói és meggyőződése alapján választotta ki, nem pedig pontosan meghatározott elvek vagy statisztikai eredmények alapján, hogy mi kerüljön bele, és mi ne kerüljön a korpuszba. Jóllehet a szakirodalomban fontos szerepet tulajdonítanak annak, hogy a korpusz reprezentatív legyen, számos nyelvészben felmerült a kérdés, hogy ez egyáltalán lehetséges-e (lásd Atkins *et al.*, 1992) különösen akkor, ha egy általános nyelvi korpuszról van szó. Erre különösen Krishnamurthy hívta fel figyelmem, rámutatva arra, hogy egyetlen nyelvre nézve sem áll rendelkezésünkre pontos statisztika, így a korpuszokat alkotó alkorpuszok százalékos aránya teljességgel önkényes (személyes közlés, Shizuoka, Japán, 2000. november 3–5). A reprezentatív korpuszokat „kiegyensúlyozott”-nak (angol: *well-balanced*) is hívja a szakirodalom, hiszen a különböző jellegű szövegek mennyisége hűen tükrözi (vagy kellene, hogy tükrözze) a mindennapi életben előfordulásuk arányát. A reprezentativitás kérdése elválaszthatatlanul összefügg a méret és a mintavétel problémáival is.

1.3.1.1. Mintavétel

Minél jobban körülhatárolható kutatásunk tárgya, annál könnyebben lehet döntéseket hozni a korpusz tartalmát illetően. Ha például egyetlen irodalmi művet kívánunk elemezni, akkor a cím megválasztásával a korpusz tartalmát is eldöntöttük. Az egy író vagy alkotó összes műveiből álló korpuszt is viszonylag könnyen elkészíthetjük. Ahogy bővítjük vizsgálatunk tárgyát, úgy szaporodnak a döntésre váró kérdések, és úgy válik megválaszolásuk is egyre nehezebbé. Ha például a regények nyelvezetét kívánjuk vizsgálni, ami még mindig meglehetősen behatárolt terület, már nem tudnánk az összes regényt, amely az adott élő nyelven megjelent, a korpuszunkba felvenni. Kénytelenek

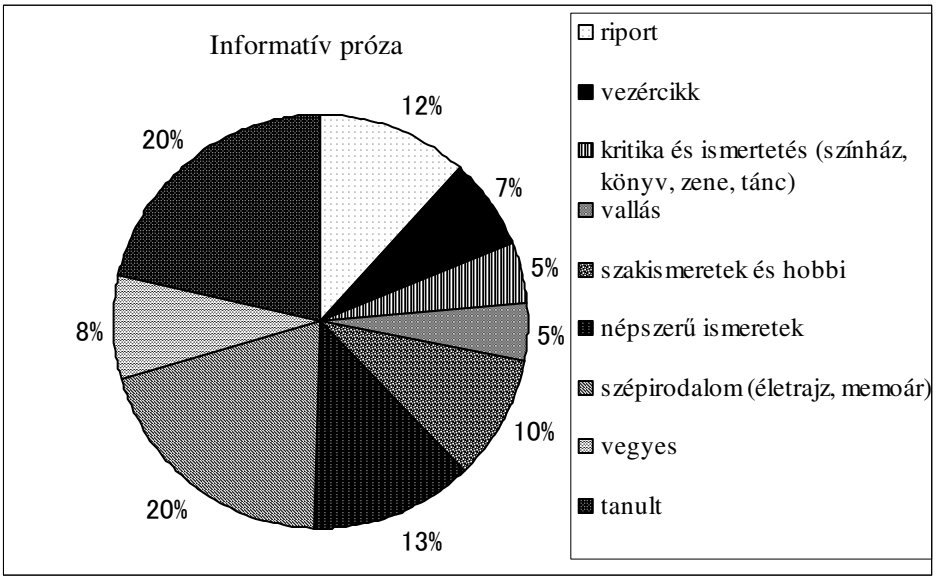
vagyunk tehát eldönteni, hogy a rendelkezésünkre álló regények közül egy adott típusú a korpusz hány százalékát tegye ki, pl. mennyi történelmi regény, sci-fi, életrajz vagy önéletrajz kerüljön a vizsgálandó gyűjteménybe. Szerepeljen-e vagy szerepelhet-e egy író összes műve, csak azért, mert könnyen elérhető? És ekkor még csak tisztán elvi kérdéseket tettünk fel, amelyeket a gyakorlatban számos más tényező, mint például a pénz, idő, és a szerzői jogok is befolyásolnak. Ebből is kiderül, hogy ha célunk maga a magyar nyelv, vagy angol nyelv, tehát egy teljes nyelv elemzéséhez szükséges korpusz összeállítása, akkor igen nagy fába vágjuk a fejszénket.

A korai korpuszok célja az volt, hogy az amerikai angol nyelvet (Brown Korpusz) és a brit angol nyelvet (Lancaster–Oslo/Bergen Korpusz (LOB) reprezentálja, így tehát mindkettőbe sok, különböző típusú szövegnek kellett bekerülnie. A 2. ábra, amely Francis és Kučera (1964) adataira épül, az első számítógépes korpusz, a Brown Korpusz fő kategóriáit, a 3. és 4. ábra az alkategóriáit mutatja be. A korpusz „informatív" prózára és szépprózára oszlik[11]. A szépprózán belüli kategóriák a következők: általános, detektívregény, tudományos-fantasztikus, kalandregény és western, romantikus és szerelmes regények, valamint humoros művek. „Informatív" próza a riport, a vezércikk, a kritika és az ismertetés, a vallási, a szaknyelvi szövegek, illetve azok, amelyek a hobbi, népszerű ismeretek, szépirodalom és „tanult" (angolul: *learned*) címszó alá sorolhatóak. Az alkategóriák további alcsoportokra oszthatók. Nézzünk meg egyetlen példát. A „tanult" kategória, amelyet a mai szóhasználatban akadémikusnak vagy tudományosnak neveznénk, 80 szöveget tartalmaz az alábbi eloszlásban: természettudomány (12), orvostudomány (5), matematika (4), társadalomtudományok (14), politikai tudományok, jog, oktatás (15), bölcsészettudományok (18) valamint technológia, mérnöki tudományok (12).

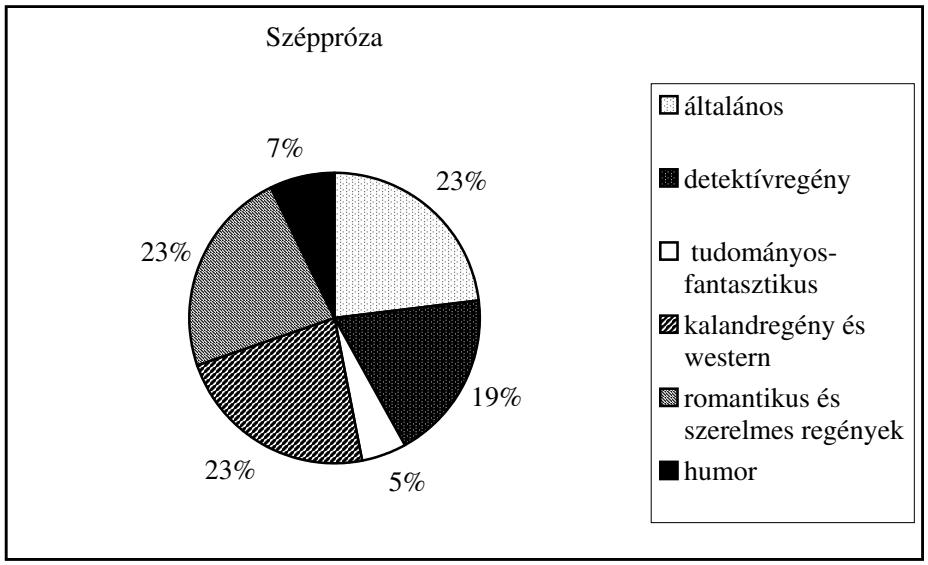


2. ábra: A Brown Korpusz összetétele

[11] Az angol terminusok esetében az *informative* az *imaginative* jelzővel áll szemben, mely a valóságra és fikcióra épülő prózát állítja szembe.



3. ábra: Az informatív próza alkategóriái a Brown Korpuszban



4. ábra: A szépprózа alkategóriái a Brown Korpuszban

Összehasonlításként nézzük meg egy reprezentatív korpusz, a Nemzetközi Angol Korpusz (International Corpus of English (ICE) http://www.ucl.ac.uk/english-usage/ice/ összeállítását, amely több alkorpuszból áll. Az egyes alkorpuszok az angol nyelv egy-egy nemzetközi változatának szövegeit tartalmazzák, és az összehasonlíthatóság érdekében mindegyik alkorpusz szerkezete egyforma. Minden szöveg kétezer szövegszóból áll, a zárójelben szereplő számok az adott csoportban szereplő szövegek számát jelentik.

BESZÉLT NYELVI (300)	**párbeszéd** (180)	**magán** (100)	beszélgetés (90) telefonbeszélgetés (10)
		nyilvános (80)	tanóra (20) közvetített beszélgetés (20) közvetített interjú (20) parlamenti vita (10) keresztkérdéses tanúkihallgatás (10) üzleti tranzakció (10)
	monológ (100)	**kézirat nélkül** (70)	kommentár (20) spontán beszéd (30) demonstráció (10) jogi képviselet (10)
		kézirattal (50)	közvetített hírek (20) közvetített beszélgetés (20) nem közvetített beszéd (10)
ÍROTT (200)	**nem nyomtatott** (50)	**nem szakmai írás** (20)	diák esszék (10) diákvizsgák kézirata (10)
		levelezés (30)	társasági levelek (15) üzleti levelek (15)
	nyomtatott (150)	**tudományos írások** (40)	bölcsészet (10) társadalomtudományok (10) természettudományok (10) technika (10)
		népszerű írások (40)	bölcsészet (10) társadalomtudományok (10) természettudományok (10) technika (10)
		riport (20)	hírjelentés (20)
		instrukciók (20)	adminisztratív írások (10) készségek/hobbi (10)
		argumentatív írások (10)	vezércikkek (10)
		kreatív írások (20)	regények (20)

2. táblázat: Az ICE összetétele a http://www.ucl.ac.uk/english-usage/ice/design.htm# honlap alapján

Az első szembetűnő különbség az, hogy ez a korpusz nem csupán írott szövegeket, hanem lejegyzett, azaz átírt beszédet is tartalmaz. Nemcsak hogy tartalmaz beszédet, de az adatok nagyobb része a beszélt nyelvből ered, nem pedig írott szövegből. Már önmagában ez a tény is jobban megfelel a valóságnak, hiszen mindennapi életünkben sokkal több beszélt nyelvi adattal találkozunk, mint írottal. Ha korpuszunkkal a nyelvet hűen akarjuk reprezentálni, akkor ezt is figyelembe kell venni. A Brown Korpuszból nem csak a beszélt nyelvi szövegek hiányoznak, hanem a nyomtatásban megjelent és meg nem jelent szövegek között sem tesz különbséget, hanem egyszerűen csak szépprózára és informatív prózára osztja őket. Pedig könnyen beláthatjuk, hogy a munkahelyen készült feljegyzés nem igazán illik egy kategóriába például a vezércikkel.

Az ICE Korpuszban az írott szövegek nyolc csoportra oszlanak funkciójuktól vagy műfajuktól függően. A különböző műfajok aránya is megváltozott. Míg a Brown Kor-

pusz szépprózа kategóriája 25%, addig az ICE korpusz „kreatív írások” kategóriája, amelyet a szépprózával gyakorlatilag azonosnak tekinthetünk, mindössze 4%-a a teljes ICE Korpusznak, és az írott szövegek alkorpuszának is mindössze 10%-a.

Természetesen a Magyar Tudományos Akadémia Nyelvtudományi Intézete is feladatának tekinti a magyar nyelv korpusz alapú leírásának elősegítését. A Korpusznyelvészeti Osztály 1997-ben alakult meg, és a következő évben kezdte meg egy akkor 100 millió szóra tervezett korpusz, a Magyar Nemzeti Szövegtár (MNSZ) http://corpus.nytud.hu/mnsz/bevezeto_hun.html készítését. Korpuszuk jelenleg 150 millió szóból áll. A szövegtár célja az, hogy *„lehetőségeihez mérten reprezentatívan tartalmazza a mai magyar nyelv jellegzetes megnyilvánulásait”* (MTA Nyelvtudományi Intézet, 1998–2002). Ezzel lehetővé válik az általános mai magyar nyelv korpusz alapú elemzése, melynek eredményei számos területen lesznek felhasználhatók. A Magyar Nemzeti Szövegtárral a 3.11. részben bővebben foglalkozunk.

A korpuszok típusairól az 1.3.2. fejezet ad átfogó képet, de már itt érdemes megemlíteni, hogy az általános és nem szakszótár jellegű kiadványok készítéséhez a lexikográfusoknak van leginkább szükségük egy általános, a teljes nyelvet reprezentáló korpuszra, hiszen az ő feladatuk az, hogy egy adott nyelvnek a lehető legteljesebb tárát készítsék el. Ideális esetben az ilyen általános korpusznak a teljes nyelv „miniatűrizált változatának” kell lennie, amely a nyelvhasználat minden aspektusáról felvilágosítással szolgál. Valóban léteznek is ilyen célból készült korpuszok, de ezek létrehozása jelentős időt, költséget, számos szakértőt igényel. Manapság, a személyi számítógépek elterjedésével egyre több az egyéni kutató, akik inkább speciális korpuszokat készítenek saját kutatási céljaikhoz, hiszen a rendelkezésükre álló forrásokból csak erre képesek. Ezek a korpuszok egy meghatározott szövegtípust tartalmaznak vagy egy meghatározott területhez kapcsolódnak, és a kutató gyakran bizonyos egyértelműen meghatározott problémának vizsgálatához készíti a korpuszt. Ezek lehetnek nyelvtani, lexikai, stilisztikai vagy diskurzus (szövegtani) elemzési problémák bizonyos szövegtípusokon belül, vagy nyelvi tankönyvek szövegei. Természetesen az ilyen korpusz szövegei csupán az adott területre nézve reprezentatívak, nem pedig az egész nyelvre. A Hongkongi Társalgási Angol Nyelv Korpusza, a Hong Kong Corpus of Conversational English (HKCCE) jó példája az ilyen speciális korpusznak. Cheng & Warren (1999) 341 kantoni anyanyelvű beszélő több mint 50 órányi beszélgetését vette fel és írta át. Ezzel egy körülbelül 500 000 szavas korpuszt hozott létre a hongkongi angol nyelvű társalgás sajátosságainak vizsgálatára. Cheng és Warren úgy vélik, hogy a társalgási nyelv teljesebb leírásával fényt deríthetnek arra a kérdésre is, hogy milyen megkülönböztető jegyek választják el más szóbeli diskurzustól a spontán párbeszédet.

1.3.1.2. A korpusz mérete

A reprezentativitás és a méret szorosan összefüggnek egymással, és jelentősen befolyásolják a kutatás hitelességét. A korpusz méretét általában a benne szereplő szavak számával adjuk meg. Viszonyítási alapként szerepeljen itt a következő két adat: egy A4-es méretű lapra kettes sorköz alkalmazásával kb. 250 szót gépelhetünk. Gárdonyi Géza *Egri csillagok* című műve pedig kb. 135 000 szóból áll. Az első elektronikusan tárolt

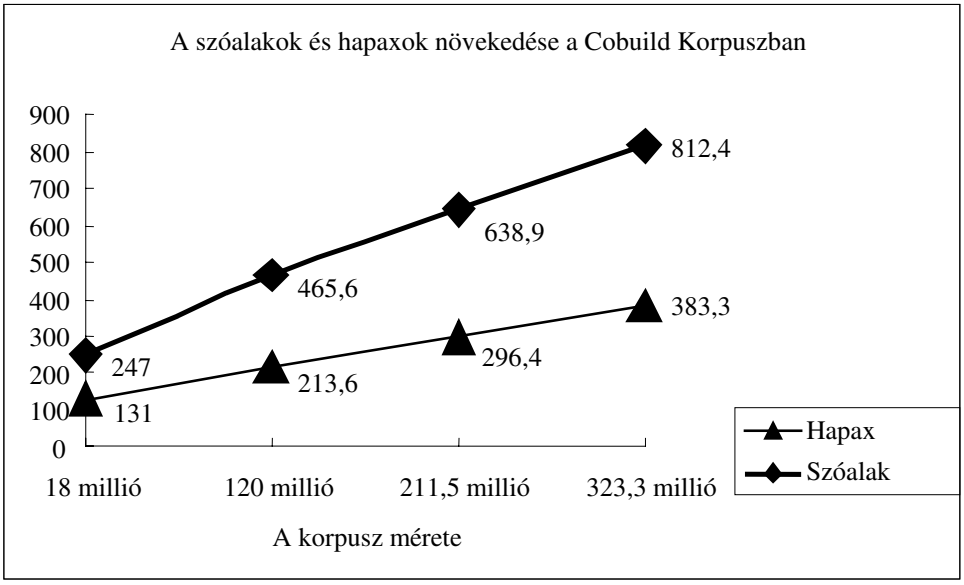
korpusz, a Brown Korpusz, az *Egri csillagok* mindössze hét és félszeresét, azaz 1 millió szót tartalmazott. A korpusz 500 darab, egyenként kb. kétezer szavas fájlból állt. Sok hasonló korpusz készült később ennek mintájára. Az angolban a méret megadásakor a szövegszó (running words) kifejezés használatos, ami a szövegben szereplő, bár többször is ismétlődő szavak számát jelenti. A számítógép számára minden, amit a jobb és a bal oldalon egy-egy szóköz határol, szónak számít. Így tehát az a mondat, hogy *„A macska felugrott a székre.”* öt szóból áll. Ha viszont azt a kérdést teszem fel, hogy hány szót kell magyarul megtanulnia valakinek, hogy ezt mondhassa, akkor a válasz csak négy, hiszen az *a* névelő kétszer is szerepel. A szakirodalomban is különbséget tesznek e kétfajta számolási mód alapján született eredmény között. Ha a szóközzel határolt szavakat számoljuk, akkor ezeket „*token*”-nek, azaz *példánynak* nevezi az angol szakirodalom, ezen tehát a korpuszban előforduló összes szót értjük, függetlenül attól, hogy ugyanaz a szó hányszor szerepel. Tehát egyszerűen megszámoljuk az összes szót a korpusz elejétől a végéig. A korpuszban szereplő különböző szavakat „*type*”-nak, azaz *szóalaknak* vagy *típusnak* nevezik. Ezen a korpuszban előforduló különböző szavak számát értjük. Nyilván egy korpuszban csak azokat a szavakat lehet vizsgálni, amelyek előfordulnak. Így könnyen belátható, hogy egy általános szótár készítéséhez olyan korpuszra van szükség, amelyben a készítendő szótár összes szava szerepel. Tehát minél nagyobb és alaposabb szótárt szeretnénk készíteni, annál nagyobb korpuszra lesz szükségünk.

Az első korpusz alapú szótár az 1980-ban megkezdett COBUILD projekt eredményeképpen látott napvilágot. A szótárkészítés megkezdésekor az eredeti korpusz mindössze 7 millió szóból állt. Jelenleg a korpusz az 500 millió szót is meghaladja. Az alábbi táblázat a COBUILD (Collins Birmingham University International Language Databank) által készített szótárakhoz használt korpusz növekedését és jellemzőit mutatja be. Érdemes megjegyeznünk, hogy az 1 millió szavas Brown Korpuszhoz képest a közel 20 évvel később készült első COBUILD Korpusz is nagy előrelépést jelentett, hiszen a hétszerese volt. Ezek után viszont csak néhány évre volt szükség ahhoz, hogy a 7 millió 18-ra növekedjen, majd a következő 5 év alatt ismét meghétszereződjön. A technika rohamos fejlődésének köszönhetően a korpuszok mérete is rohamosan nőtt, amit az alábbi táblázat mutat be.

1.	COBUILD mérete	18 millió szó	120 millió szó	211 millió szó	323 millió szó
	Időpont	1987	1993	1995. április	1996. július
2.	Összes szó Példányok száma	18 000 000	120 000 000	211 505 963	323 302 789
3.	Különböző szavak száma: Szóalakok/típusok	247 069	475 633	638 901	812 452
4.	Csak EGYSZER előforduló szavak: Hapax legomena	131 299	213 684	296 436	383 356
5.	TÖBBSZÖR előforduló szavak: Nem hapax	115 770	269 949	342 464	429 096
6.	10-nél többször előforduló szavak:	43 579	104 201	134 942	164 963
7.	15-nél többször előforduló szavak:	nincs adat	nincs adat	111 007	164 633

3. táblázat: A korpusz mérete és jellemzői Krishnamurthy (1997b) alapján

A táblázatban az eddig előfordult kifejezések mellett szerepel még a *hapax legomena* kifejezés, ami a görög *hapax legomenon*, azaz „egyszer mondott” többes számú alakja. Ahhoz, hogy egy szót a szövegkörnyezetében megvizsgáljunk, általában nem elég, ha csak egyszer találkozunk vele. Márpedig a fenti táblázatból az derül ki, hogy bármilyen nagy is a korpusz, az egyszer előforduló szavak az összes különböző szónak (típusnak) majdnem felét alkotják. (Lásd a táblázat 3. és 4. sora.) A 10-nél többször előforduló szavak csak a teljes korpusz töredékét teszik ki.



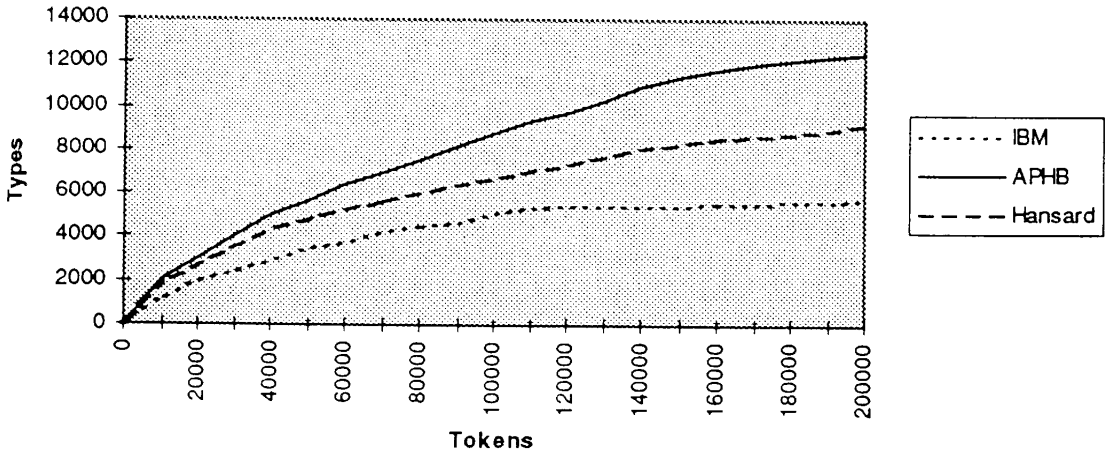
5. ábra: A típusok és a hapaxok növekedése a COBUILD Korpuszban

A fenti ábrából világossá válik, hogy ha a 18 milliós korpuszt több mint 16-szorosára bővítjük, mindössze háromszoros a benne levő szavak, típusok száma, s a 10-nél többször előforduló szavak is csak négyszeresükre nőnek. Így könnyen belátható, hogy a korpusz jelentős növelése is csak szerény növekedést jelent a típusok növekedésében és a típusok előfordulásának megnövelésében. Sinclair szerint *„a korpusznak a lehető legnagyobbnak kell lennie … 10 példa szegényes minta; legalább 50-re van szükség, hogy egy szó jelentéseit körvonalazhassuk, és 150-re van szükség ahhoz, hogy megbízhatóan számoljunk be róluk”*[12] (1993: 7).

A másik probléma az, hogy a típusok száma ebben a formában mindent magában foglal: személyneveket, helységneveket, gépelési hibákat, amelyek látszólag új szavakat (típusokat) eredményeznek. Ezeket természetesen nem lehet a lexikográfiai vizsgálathoz használni. Így az előbb említett számok csak egy kalkulált, nem pedig tényleges növekedést mutatnak.

McEnery és Wilson (1996: 157) három különböző korpusz, a Hansard, az IBM és az American Printing House for the Blind (APHB) korpusz növekedésére vonatkozóan ad meg hasonló adatokat.

[12] Eredetiben: “*Corpora should be as large as possible…ten is a poor sample; you need 50 at least to appreciate the meaning profile of a word, and 150 for any reliable account of its meaning…*” (1993: 7).



6. ábra: Az IBM, APHB és a Hansard Korpusz lexikonjának növekedése (McEnery és Wilson (1996: 157)

Itt kell szólnunk még egy fontos fogalomról, a lemmatizálásról és a lemmákról. Jóllehet a következő szavak formája különböző: *eszem, eszik, ettetek*, mégis a hétköznapi életben is egy szónak tekintjük ezeket, hisz ugyanannak a szótári egységnek a ragozott változatai. Ha a korpuszban az ilyen alakokat egy csoportba vonjuk, azzal máris több előfordulását vizsgálhatjuk egy bizonyos szónak, azaz lemmának. Léteznek olyan számítógépes programok, amelyek ezt automatikusan elvégzik. Természetesen a magyar nyelvben az angolnál sokkal több különböző formával találkozhatunk, így a lemmatizálás még fontosabb a magyar nyelv és minden morfológiailag gazdag nyelv esetében.

A korpuszok – különösen a korai, kisebb korpuszok – sok esetben nem a teljes szöveget tartalmazzák, hanem csak egy töredékét minden szövegnek. Ez például jelentheti azt, hogy cikkek esetében a konklúzió soha nem kerül be a korpuszba vagy csak ritkán. Ennek egyenes következménye az, hogy a tipikusan a szöveg végén megjelenő szavak és kifejezések sokkal kisebb számban fordulnak majd elő a korpuszban, mint a tipikusan a bevezetésben használtak. Még egy nagy hátránya az ilyen „csonka” szövegekből álló korpusznak, hogy a nagyobb szövegszerkezeti jellemzőket nem vizsgálhatjuk a segítségükkel (Biber & Finegan, 1994). Éppen ezért érvelt Sinclair (1996a) is amellett, hogy a nagyméretű korpuszokat lehetőség szerint teljes szövegekből kell létrehozni. Ezzel szemben néhányan úgy vélik, hogy bizonyos nyelvtani minták vagy kifejezések vizsgálatához nem feltétlen szükséges, hogy nagy korpusszal rendelkezzünk. Kjellmer úgy véli, hogy jóllehet a kisméretű korpuszok, mint a mindössze 1 millió szavas Brown és a LOB Korpusz, vagy az 500 000 szavas London-Lund Korpusz túl kis méretűek a lexika vizsgálatához, de elegendő nagyságúak ahhoz, hogy az angol nyelv kifejezéseinek jelentős részét mai használatukban megtaláljuk bennük. Kjellmer (1991: 115) írja:

> Szerencsés dolog tehát, hogy a behatároltabb korpuszok, mint az 1 millió szavas amerikai *Brown Korpusz* vagy brit megfelelője, a *LOB Korpusz*, sőt még a *London-Lund Beszélt Nyelvi Angol Korpusz*, a maga 500 000 szavával annak ellenére, hogy a lexika vizsgálatá-

> hoz kissé kevés lenne, mégis elegendők ahhoz, hogy a jelenleg használt angol kifejezések jelentős része szerepeljen bennük[13]

Jóllehet sok „előgyártott" korpusz létezik elsősorban az angol nyelvet vizsgálók számára (Brown, LOB, British National Corpus, COBUILD Direct stb.), egyre több tanár és kutató érzi szükségét annak, hogy saját maga hozzon létre korpuszt, hiszen a „kész" korpuszok nem mindig felelnek meg tanítás vagy saját kutatás céljaira. A magyar nyelvvel foglalkozók kénytelenek elkészíteni saját korpuszukat, hiszen nem létezik mindenki számára könnyen hozzáférhető magyar nyelvű korpusz, még akkor sem, ha az interneten már rendelkezésre áll egy kereső, amellyel a Magyar Nemzeti Szövegtárban kereshet bárki, aki az ingyenes regisztrációt elvégzi. Természetesen a „házi készítésű" korpuszok nem vehetik fel a versenyt olyan korpuszokkal, amelyeket nagy cégek, egyetemek és állami kutatási alapok is támogatnak. Nem is feladatuk, hiszen nem az egész nyelvet kívánják elemezni, hanem valamilyen jól megfogható probléma vizsgálatára készülnek. Ez lehet nyelvtani, lexikai, stilisztikai, vagy az adott nyelvet tanulók problémáinak vizsgálata.

Mivel az interneten található ilyen jellegű korpuszok száma napról napra gyarapszik, a teljesség igénye nélkül szeretném felhívni a figyelmet néhányra közülük. A CANCODE (Cambridge and Nottingham Corpus of Discourse in English) több szempontból is figyelemre méltó. Egyrészt azért, mert az ezen a korpuszon alapuló kutatások eredményei megtalálhatók folyóiratokban és könyvekben, valamint számos konferencián is beszámoltak róla, elsősorban a korpusz létrehozói, azaz Michael McCarthy és Ronald Carter. Másrészt a korpusz beszélt angol nyelvet tartalmaz, és ilyen korpuszból kevés van. Harmadrészt azért említésre méltó, mert számos cikkben azt tárgyalják a szerzők, hogy a kutatás eredményeit miként lehet a nyelvtanításban felhasználni.

Szeretnék még egy nagyszabású egyéni vállalkozásról beszámolni, amelyet Szakos József végez Tajvanon. A tajvani őshonos tsou, saaura és kanakanavu nyelveket a kihalás veszélye fenyegeti. Ezek a nyelvek még írott változattal sem rendelkeznek, így még nehezebb a korpusz létrehozása. Gyakorlatilag a munka azzal kezdődik, hogy Szakos magnóval felfegyverkezve felkeresi az ezen nyelveket beszélők falvait, felveszi beszédüket, majd hazatérve a felvétel szövegét átírja a számítógépen. Érthető tehát, hogy ez a korpusz folyamatosan, és igen lassan készül. 2001. januárjában a tsou korpusz 400 000 szóból, a saarua 100 000 szóból, a kanakanavu pedig 50 000 szóból állt.

Ha alaposabban belegondolunk, a 400 000 szavas tsou korpusz majdnem fele a Brown Korpusznak, amely sokáig az egyetlen angol nyelvű korpusz volt, és amely kutatásra ma is az egyik legtöbbet használt korpusz. Ha 1 millió szó alkalmas volt az angol nyelv kutatására, talán a 400 000 szavas tsou korpusz is nagy előrelépést jelent egy írásbeliséggel nem rendelkező nyelv leírásában.

---

[13] Eredetiben: "*It is fortunate, then, that more limited corpora, like the one-million word American Brown Corpus or its British counterpart, the* LOB Corpus, *and even the* London-Lund Corpus of Spoken English *with about 500,000 words, although they are on the small side for a study of lexis, are none the less large enough to contain a considerable part of the English phrases in current use"* (1991: 115).

A korpusz méretének tárgyalását szeretném Sinclairt idézve lezárni, aki a mai napig is azt vallja, hogy minél több adatra, minél nagyobb korpuszra van szükség. Sinclair (2001) arra a következtetésre jut, hogy a két fajta korpusz különböző megközelítést igényel, az ő elnevezésével „korai emberi beavatkozás"-t *(early human intervention)* és „késleltetett emberi beavatkozás"-t *(delayed human intervention).* A korai beavatkozás a kis korpuszok esetében használatos. Ez azt jelenti, hogy miután a korpusz a számítógép segítségével elkészül, a nyelvi elemzés kézi munkával történik, s esetleg követheti egy ismételt számítógépes alkalmazás. A nagy korpuszok esetében a gépi elemzés több program futtatásával kezdődik. Ezek a programok is kézi elemzés eredményeképpen jöttek esetleg létre, de nem a teljes korpuszon végezték a kézi elemzést, hanem egy kisebb korpuszon. Az eredmények végső értelmezése mindenképpen humán beavatkozást igényel, de a kis korpuszokhoz képest ez sokkal később történik. Ezért ezt késleltetett beavatkozásnak nevezi.

1.3.2. A korpuszok fajtái

1.3.2.1. A mintavétel módja szerint

Statikus korpusz
A korpusz tervezésekor el kell dönteni, hogy mi kerül bele és mi nem. Azt is érdemes előre eldönteni, hogy később kívánunk-e rajta változtatni, akarjuk-e bővíteni vagy sem. A Brown, a LOB és az összes mintájukra készült 1 millió szavas korpusz változatlan. A nyelvet ezek a korpuszok egy bizonyos időpontban, mintegy pillanatfelvételként ábrázolják, így tehát alkalmasak arra, hogy összehasonlító vizsgálatokat végezzünk velük. Természetesen egy sokkal nagyobb méretű korpuszt is lehet statikusra tervezni.

Dinamikus korpusz
A legismertebb dinamikus korpusz a Cobuild Korpusz. Folyamatosan bővítik, és jóllehet néhány fájl törlésre kerül, a fő cél az állandó növekedés. A rendszeres növelésnél az arányok nem állandóak.

Monitor korpusz
Sinclair (1991) egy monitor korpusz létrehozását is megemlíti, amely mintegy kombinációja a statikus és a dinamikus korpusznak. Az eredeti (statikus) korpuszhoz bizonyos időközönként – havonta, évente – új szövegeket adnak hozzá, de az eredeti korpusz arányait mindig megtartva. Így a dinamikus, új adatokat tartalmazó alkorpusz adatai összehasonlíthatók az eredetivel, vagy bármely más időpontban hozzáadott alkorpusszal, amely lehetővé teszi, hogy a nyelv változását nyomon kövessük. Monitor korpuszról mostanában nemigen hallani, pedig sok szempontból hasznos információkkal szolgálhatna.

1.3.2.2. A korpusz felhasználásának módja szerint

A mintavétel, reprezentativitás és a méret tárgyalásakor már említést tettünk néhány jellemzőről, amelyeket most más csoportosításban ismét megemlítünk.

Általános korpusz

Az általános korpusz készítésének legfőbb célja egy adott nyelv minél hitelesebben történő reprezentálása. Nyilvánvaló, hogy lehetetlen egy nyelv összes írott és szóbeli megnyilvánulását számba venni. Azt is igen nehéz eldönteni, hogy egy nyelv használatában milyen arányban szerepnek a különböző műfajok. Lehetséges-e pontosan megválaszolni még akár azt az egyszerű kérdést is, hogy valójában milyen a beszélt nyelv és az írott nyelv aránya? 60% : 40%? Vagy 70% : 30%? Nem igazán. Ezért mindössze azt mondhatjuk, hogy törekedni kell a becsléseken alapuló, általánosan helyesnek ítélt arányok betartására. Az is könnyen belátható, hogy mivel minden műfajnak szerepelnie kell egy ilyen korpuszban, és megfelelő mennyiségre is szükség van a nyelvi elemzés elvégzéséhez, az általános korpusz méretének minél nagyobbnak kell lennie.

Az általános korpuszra elsősorban a lexikográfusoknak van szükségük. A korpusz alapján vizsgálják a szavak használatát, jelentését és az ezekben bekövetkező változásokat. Nyelvtanok, nyelvleírások és más általános referencia jellegű művek is ilyen korpuszok elemzésével készülnek. Az általános korpuszok viszonyítási alapként is használhatók, hiszen a speciális céllal készült korpuszok jellemzőire csak akkor derülhet fény, ha egy általános korpusszal össze lehet őket hasonlítani.

Példaként említhetjük a következőket: Eredetileg a Brown és a LOB Korpusz is ilyen általános céllal készült, de ma már a méretük miatt nem szolgálják megfelelően az eredeti célt. A modern nagy korpuszok közül megemlíthetjük a Bank of English-t (legutóbbi adat szerint 524 millió szó), amelyre gyakran COBUILD Korpuszként is utalnak. A Brit Nemzeti Korpusz (British National Corpus, azaz BNC, 100 millió szó) szintén általános korpusz, amelyre a nemzeti jelző alapján következtethetünk is. Mint azt az 1.3.1.1. pontban is említettük, a Magyar Nemzeti Szövegtár (MNSZ) is ezzel a céllal készült, és a tervbe vett 100 millió szó helyett már 153,7 millió szövegszót tartalmaz a honlap adatai szerint (2005. május).

Speciális korpuszok

Tulajdonképpen minden, az általánostól eltérő korpusz speciálisnak nevezhető. Az ilyen korpuszok készítésekor a vizsgálat tárgyának és céljának megfelelően kell a korpuszba kerülő szövegeket kiválasztani. Az ilyen korpusz készítésekor a cél lehet egy műfaj, vagy egy társadalmi réteg nyelvezetének vizsgálata, de más szempontok is szolgálhatnak a kiválasztás alapjául. A specializáció bármilyen mértékű lehet, pl. az orvos és beteg közötti párbeszéd, vagy akár külön-külön vizsgálhatjuk az orvosi beavatkozás előtti és utáni párbeszédet. Az orvosi beavatkozás alatti, pl. operáció közbeni, orvos és asszisztensek közötti kommunikációt is vizsgálhatjuk. Pedagógiai szempontból hasznos lehet egy tanár-diák közötti, tanórai vagy azon kívüli korpusz összeállítása és vizsgálata is.

Példák: Az 1.3.1.1. rész végén említett Hongkongi Társalgási Angol Nyelv Korpusza (Hong Kong Corpus of Conversational English) (HKCCE) jó példája az ilyen speciális korpusznak. A Cambridge and Nottingham Corpus of Discourse in English (CANCODE), (5 millió szó) az informális brit angol vizsgálatához készült.

Összehasonlítható korpuszok (*comparable corpora*)

Akár általános, akár speciális korpuszról van szó, ha azonos szempontok szerint állították össze őket, és a méretük is azonos, akkor összehasonlíthatók. Számos korai korpusz azzal a céllal követte az elsőként összeállított Brown Korpusz mintáját, hogy megteremtse az összehasonlítás lehetőségét, ezért mind a méretük, mind a mintavétel elvei megegyeznek. Ezek a „Brown klónok" a következők: a LOB**,** Kolhapur Corpus of Indian English (KOL), a Freiburg Korpusz**,** az Australian Corpus of English (ACE), amely Macquarie Corpus of Written Australian English néven is ismeretes, valamint a Wellington Corpus of Written New Zealand English**,** az International Corpus of English (ICE) és a Corpus of English-Canadian Writing.

Nem csak egy nyelv különböző változatait lehet összehasonlítani, hanem két vagy több különböző nyelv korpuszát is, ha azok azonos szempontok alapján készültek (azonos mennyiségű szöveget tartalmaznak azonos műfajokon belül). Az ilyen többnyelvű korpuszt a fordítók és nyelvtanulók is egyaránt jól használhatják.

Párhuzamos korpusz (*parallel corpus*)

Az előbb említett két vagy több nyelvű összehasonlítható korpusztól abban különbözik, hogy a párhuzamos korpuszban azonos szövegek különböző nyelvű fordításai szerepelnek. Ez lehet egy magyar regény és angol fordítása, valamint egy angol regény magyar fordításából álló korpusz. Az ilyen korpusz a fordítókat és a nyelvtanulókat segíti az egyes kifejezések fordítási megfelelőinek azonosításában és vizsgálatában. Jóllehet fordítási korpuszként (*translation corpus*) is utalnak erre a korpuszra, helyesebb, ha ezt a kifejezést kizárólag a csak fordításokból álló, egynyelvű, eredeti műveket nem tartalmazó korpuszra használjuk. Például francia regények magyar nyelvű fordítását tartalmazza. A párhuzamos korpusz esetében két vagy több különböző nyelvű, de azonos szöveg a számítógép képernyőjén egymás mellett látható. Az eredeti mondat mellett látható a fordítása vagy fordításai, és innen származik a párhuzamos elnevezés. Példa erre: English-Swedish Parallel Corpus (Svéd–angol Párhuzamos Korpusz), mely körülbelül kétszer 1 millió szóból áll (Altenberg & Aijmer, 2000).

Nyelvtanulói korpusz

Egy bizonyos nyelvet idegen nyelvként tanulók által létrehozott szövegek gyűjteménye. Természetesen tartalmazhat szóbeli megnyilvánulásokat is, nem csak írott szövegeket. Az ilyen korpusz azzal a céllal készül, hogy fényt derítsen arra a kérdésre, hogy miben különbözik a tanulók nyelvezete az anyanyelvi beszélőkétől. Érdekes eredményeket hozhat a különböző anyanyelvű diákok idegen nyelvű megnyilvánulásainak összehasonlítása. Sok gyakorló pedagógus készít diákjai munkáiból álló korpuszt, de ezek viszonylag kis méretűek, és sokszor csak az általános korpusszal lehet őket összehasonlítani. Példa: A legismertebb korpusz az International Corpus of Learner English (ICLE). Ebben a korpuszban az egyes alkorpuszok különböző anyanyelvű diákok (francia, német, magyar stb.) munkáit tartalmazzák, és egyenként 200 000 szóból állnak. Magyar példa is van: a Pécsi Tudományegyetem angol szakos hallgatóinak esszéiből készített Horváth József (2000) egy 412 000 szóból álló korpuszt pedagógiai céllal. Az egyik legnagyobb azonos anyanyelvű diákok munkájából készült korpusz a Hong Kong

University of Science and Technology Corpus (HKUST), amely kínai anyanyelvi beszélők angol nyelvű munkáit tartalmazza.

Pedagógiai korpusz

Elsőként Dave Willis (1990) írt ilyen korpuszról, de ő akkor „tanulói" korpusznak nevezte. Eredeti értelmezése szerint olyan szövegek gyűjteménye, amelyekkel a nyelvtanuló tanulmányai során szembesül. Ez egyben azt is jelenti, hogy minden egyes tanuló, még ugyanazon csoportba tartozó tanulók esetében is, különböző korpusszal rendelkezik, és a tanár vagy bárki más számára ez hozzáférhetetlen és megismételhetetlen. Ezért ebben az értelmezésben nem használható gyakorlati kategóriaként. Willis jelenlegi, módosított értelmezése szerint a pedagógiai korpusz azon szövegek összessége, amelyekkel egy adott kurzuson a résztvevő diákok találkoznak. Tulajdonképpen ez a kurzus teljes anyagát jelenti. Ha a tanfolyam anyaga autentikus szövegekből áll, akkor ez a pedagógiai korpusz is autentikus. Az ilyen korpusz számos előnnyel rendelkezik. A legtöbb korpusz esetében a vizsgált szavak és kifejezések bővebb kontextusa ismeretlen, még akkor is, ha egy több sorból álló részletet „előhívhatunk". A diákok a pedagógiai korpusz esetében a vizsgált szavak bővebb kontextusára is emlékeznek, hiszen már olvasták vagy hallották a szöveget. Ez a korpusz a már tanult anyagok új szempontok szerinti felfedezését teszi lehetővé, a diákok a tanár „beavatkozása" nélkül kaphatnak választ kérdéseikre, vagy fedezhetik fel a nyelv szabályait. Köztudott, hogy az ismétlésnek, a tudás rendszerezésének milyen jelentős szerepe van a nyelvtanulásban. E korpusz használatával a diákok szinte játékosan ismételhetnek. Az ilyen korpuszok a diákok nyelvi szintjének is megfelelnek, így a tanár „közvetítése" nélkül, a diákok önállóan is használhatják. A tanárok számára is hasznos „alapanyagként" szolgálhat feladatok összeállításához.

Történeti vagy diakrón korpusz

Az adott nyelv történeti változásainak követéséhez, a múltbeli adatok feldolgozásához összeállított korpusz. A különböző időszakokból származó szövegek gyűjteménye lehetővé teszi a nyelv változásának követését és elemzését. A legismertebb ilyen jellegű korpusz az International Computer Archive of Modern and Medieval English, mely ICAME (ejtsd: ájkéim) néven ismeretes. A honlapja http://helmer.hit.uib.no/icame.html címen érhető el. A Penn–Helsinki Parsed Corpus of Middle English nagyrészt az ICAME Középkori Angol Alkorpuszára épült, és elsősorban a mondattan vizsgálatában használható. 1,3 millió szövegszóból álló prózai szövegek gyűjteménye.

A Tycho Brahe Parsed Corpus of Historical Portuguese portugál írók 1500 és 1900 között született műveiből álló gyűjtemény, melyet 1 millió szóra terveztek. A korpusz segítségével a modern európai portugál nyelv kialakulását elemezhetik és követhetik nyomon.

A Magyar Tudományos Akadémia Nyelvtudományi Intézetének Lexikográfiai és Lexikológiai Osztálya készítette a Magyar Történeti Korpuszt, melynek honlapja a következő címen található: http://www.nytud.hu/adatb/index.html. A korpusz 23 millió szövegszóból áll, és a http://www.nytud.hu/hhc/ lapon található a mindenki számára hozzáférhető kereső.

1.3.3. Jogi problémák

Jóllehet a szerzői jogra vonatkozó törvények országonként igen eltérőek lehetnek, léteznek nemzetközileg elfogadott szabályok. Néhány dologra szeretném csak felhívni a figyelmet, nem pedig kimerítő információval szolgálni. Alapvetően akár fénymásolatról, akár elektromos másolatról van szó, a szerzői jog tulajdonosának előzetes írásbeli beleegyezése nélkül jogellenes mind fénymásolatot, mind pedig elektromos másolatot készíteni. Manapság ez nem csak teljes művekre, cikkekre, hanem részletekre is vonatkozik. (Magyarországon is az Európai Unióban általánosan elterjedt szabályozás érvényes: az írásművek a szerző halála után 70 évvel válnak szabadon felhasználhatóvá. Ezt megelőzően az írásmű felhasználásához a jogtulajdonos engedélye szükséges.) A felhasználás célja szerint is különbözhetnek a vonatkozó rendelkezések. Oktatási vagy saját célokra való felhasználás esetén sokszor engedélyezett a másolat készítése, de minden esetben érdemes jogi tanácsot kérni, mielőtt korpusz készítéséhez kezdünk. Ha szükséges a szerzői jog tulajdonosától engedélyt kérni, akkor ez komoly feladatot jelenthet egy magánszemély részére.

Nem véletlen azonban, hogy az interneten a legtöbb klasszikus művet megtalálhatjuk letölthető formában, hiszen a mai értelemben vett szerzői jogok vagy soha nem is léteztek e művek esetében, vagy pedig elévültek. Shakespeare, Dante, a Biblia különböző változatai több nyelven is megtalálhatók. A magyar nyelv esetében a Magyar Elektronikus Könyvtárból http://www.mek.iif.hu/porta/ lehet dokumentumokat, teljes műveket letölteni, de minden egyes dokumentum esetében a következő figyelmeztetéssel találkozunk:

```
//////////////// MAGYAR ELEKTRONIKUS KÖNYVTÁR \\\\\\\\\\\\\\\\

Ez a dokumentum a Magyar Elektronikus Könyvtárból származik. A
szerzői és egyéb jogok a dokumentum szerzőjét/tulajdonosát
illetik (amennyiben az illető fel van tüntetve). Ha a szerző
vagy tulajdonos külön is rendelkezik a szövegben a terjesztési
és felhasználási jogokról, akkor az ő megkötései felülbírálják
az alábbi megjegyzéseket. Ugyancsak ő a felelős azért, hogy
ennek a dokumentumnak elektronikus formában való terjesztése
nem sérti mások szerzői jogait. A MEK üzemeltetői fenntartják
maguknak a jogot, hogy ha kétség merül fel a dokumentum szabad
terjesztésének lehetőségét illetően, akkor töröljék azt a MEK
állományából.

Ez a dokumentum elektronikus formában szabadon másolható,
terjeszthető, de csak saját célokra, nem-kereskedelmi jellegű
alkalmazásokhoz, változtatások nélkül és a forrásra való
megfelelő hivatkozással használható. Minden más terjesztési és
felhasználási forma esetében a szerző/tulajdonos engedélyét
kell kérni. Ennek a copyright szövegnek a dokumentumban mindig
benne kell maradnia.

A Magyar Elektronikus Könyvtár elsősorban az oktatási/kutatási
szférát szeretné ellátni magyar vagy magyar vonatkozású,
```

```
szabad terjesztésű elektronikus szövegekkel. A MEK projekttel
kapcsolatban a MEK-L@listserv.iif.hu lista e-mail címén lehet
információkat kapni és kérdéseket feltenni. A MEK központi
Internet szolgáltatásainak URL címei: <http://www.mek.iif.hu>,
<gopher://gopher.mek.iif.hu> és <ftp://ftp.iif.hu/pub/MEK>.

\\\\\\\\\\\\\\\\ HUNGARIAN ELECTRONIC LIBRARY ////////////////
```

7. ábra: A Magyar Elektronikus Könyvtár figyelmeztető szövege

Legutóbbi internetes keresésünk során a következő címen is elértük a fenti adatbázist: http://mek.oszk.hu/katalog/. Az itt szereplő katalógus kezelőfelülete igen jól használható, és a megtekintéskor vagy letöltéskor választható fájlok száma nőtt, s általában html, word, rtf és pdf formátum között választhatunk. A fenti szerzői jogi figyelmeztetés természetesen itt is azonos szövegű.

1.3.4. Átírás és annotáció

1.3.4.1. A beszéd átírása

A legtöbb korpusz írott szövegekből áll, kevés az olyan, amely élő beszéd rögzítését is tartalmazza. Ha mégis van beszélt nyelvi összetevője, a legtöbb akkor is csak a szavakat tartalmazza, azaz átírt szöveg, mely az élő szó zenei eszközeire csupán írásjelekkel utal. Tudjuk azonban, hogy a leírt szöveget sokféle intonációval lehet felolvasni, és arra vonatkozóan, hogy mely esetben mi hogyan hangzott el, az ilyen fajta átírás semmi információval nem szolgál. Létezik azonban néhány olyan speciális korpusz, amelynek az is célja, hogy a beszédre jellemző egyéb jegyeket is a lehető legpontosabban visszaadja. Példa erre a Lancaster–IBM Spoken English Corpus, amelyben 15 speciális karakter segítségével jelölik a tonetikus hangsúlyt, azaz a hangszín és hangmagasság változását (Knowles *et al.*, 1996: 61).

012 SPOKEN ENGLISH CORPUS TEXT A12
From our own Correspondent
Broadcast notes: Radio 4, 11.30 a.m., 22nd June, 1985
Transcribed by BJW

¯there have been ˏtimes | in this ¯past few ˬweeks | when ↑¯what's been _happening in Hong ˇKong has been | like ˋone of those ¯old _old _western ˋmovies || ˇyou know | the ˋones where | ↑the _wagons are ·gathered into a ˋtight circle on the ˋhostile ˇprairie | ˋunder attack and ˋdown to the last ˇbullet | when ¯lo and be_hold | the ˋUS ˇcavalry | ˋrides over the ·hill to the ˋrescue || ¯well it's _been | ↑the _government ˋbankers and ˋstock ex-

8. ábra: Prozódikus átirat

Az ilyen átírás nem csak hogy időigényes, de az átírók megfelelő képzettsége nélkül teljesen elképzelhetetlen, hiszen nagyon sok múlik az átírás következetességén. Az átírás szinte minden problémáját tárgyalja a Leech, Myers & Thomas (1995) által szerkesztett *Spoken English on computer: Transcription, mark-up and application* ['Beszélt angol nyelv a számítógépen: Átírás, jelölés és alkalmazás'] című könyv, mely 20 cikket tartalmaz. Cook (1995) és Edwards (1995) az átírás általános és elméleti problémáival foglalkozik, míg Sinclair (1995) az elmélet gyakorlati alkalmazásáról ír. De nem csak általános problémákat érintő művek léteznek (vö. Cheepen, 1995; Perkins, 1995; Roach & Arnfield, 1995; Sebba, 1995), hanem az egyes korpuszokra vonatkozó problémákat taglaló művek is: A Brit Nemzeti Korpuszra (BNC) vonatkozik Crowdy (1995) és Garside (1995); a Londoni Tinédzser Nyelv Bergeni Korpuszára (Bergen Corpus of London Teenager Language, COLT) vonatkozik Haslerud & Stenström (1995); a Nemzetközi Angol Korpuszról (International Corpus of English) ír Nelson (1995); a COBUILD Élő Beszéd Korpuszáról (COBUILD Spoken Corpus) Payne (1995), az Angol Nyelvhasználati Felmérés (Survey of English Usage, SEU) és a London–Lund Beszélt Nyelvi Korpuszról (Corpus of Spoken English (LLC) Peppé (1995); és a Humán Kommunikációs Kutatási Központ Térkép Feladatának Korpuszáról (Human Communication Research Centre's Map Task Corpus) Thompson, Anderson & Bader (1995). A felsorolás csak példaértékű, nem a teljesség igényével készült.

A modern technológia fejlődése azonban lehet, hogy hamarosan lehetővé teszi, hogy a hangfelismerés (*voice recognition*) segítségével a jövőben a szöveg leírása nélkül lehessen a rögzített adatokat elemezni. Az is elképzelhető, hogy a szöveg átírása teljesen automatizált legyen, és minden emberi beavatkozás nélkül (vagy csak minimális beavatkozással) készüljön. Amíg ez nem következik be, valószínűleg nem fogunk bővelkedni beszélt nyelvi korpuszokban, hiszen mind a szakemberek képzése, mind pedig az átírási folyamat rengeteg időt és anyagi fedezetet igényel.

1.3.4.2. A standard annotáció

Korpuszannotációnak nevezünk minden olyan információt és jelet, amelyet az eredeti szöveg (akár írott akár beszélt nyelvi) nem tartalmazott, de a korpusz készítésekor vagy feldolgozása során a szövegbe bekerült. Ez a „plusz" információ nagyon megkönnyíti az adatok lekérdezésének a gyorsaságát és pontosságát, valamint a szöveg azonosítását. Lássunk erre egy egyszerű példát. Ha olyan szót vizsgálunk egy korpuszban, amely homonimával (azonos alakú, de teljesen különböző jelentésű szóval) rendelkezik, pl. *ég, vár,* melyek egyaránt lehetnek igék vagy főnevek, akkor az eredmény nem csak a keresett szavakat tartalmazza, hanem annak homonimáit is. Ha a korpusz nagy méretű, ez esetleg több ezer vagy tízezer adat kézi ellenőrzését tenné szükségessé minden egyes lekérdezés alkalmával. Ha azonban a korpuszban már minden szót szófajilag megjelöltünk, a lekérdezés során csak a kívánt szófajú adatok jelennek meg, melyek még ekkor is tartalmazhatnak azonos szófajú homonimákat, pl. *ár*, mint szerszám vagy fizetendő összeg. Természetesen, ha az annotáció nem pontos, akkor az ezt felhasználó kutatás sem lehet az.

A korpuszban leggyakrabban előforduló nyelvi annotáció a szófaji címkézés *(tagging*), vagy azonosítás, és a szintaktikai (mondattani) kapcsolatokat azonosító elemzés (*parsing*). A szófaji címkézés esetében a szöveg minden egyes szava mellett szerepel a

szófaji megjelölés is. Egy mindössze 1 millió szóból álló korpusz címkézésének kézi végrehajtásához is rengeteg időre és képzett szakemberre van szükség, így a manapság megszokott 100 millió szó kézi címkézése teljesen lehetetlen feladat lenne, de nincs is már erre szükség. Egy kézi címkézéssel készült kisebb korpuszt használnak a számítógép „betanítására", és a program tökéletesítése után a címkézés automatikusan, a gép ellenőrzése nélkül történik. A címkézés a kutatók munkáját jelentősen tehermentesíti, hiszen csak az általuk valóban vizsgálni kívánt adatok jelennek meg a keresés végeredményeként. Így nem kell napokat vagy heteket tölteni a „rostálással". Az MNSZ a következő alapkódokat használja (forrás: http://corpus.nytud.hu/mnsz/sugo_hun.html #szofajkod):

N	főnév	*DIG*	számjeggyel írt szám
A	melléknév	*Det*	névelő
Num	számnév	*N*	névutó
MIA	beálló melléknévi igenév	*V*	ige
MIB	befejezett melléknévi igenév	*Pre*	igekötő
MIF	folyamatos melléknévi igenév	*V.INF*	főnévi igenév
Pro	névmás	*V.HIN*	határozói igenév
Adv	határozószó	*Con*	kötőszó
Int	indulatszó	*ELO*	előtag
S	mondatszó	*WPUNCT*	központozás
Abb	rövidítés	*SPUNCT*	mondatvégi írásjel

4. táblázat: A Magyar Nemzeti Szövegtár alapkódjai

Természetesen ezek az alapvető kategóriák nem elegendők a pontos kereséshez, mert a melléknevek esetében pl. a fokozott alakokra lehetünk kíváncsiak, vagy az igék esetében csak bizonyos személyű vagy számú alakokra, a főnevek esetében pedig a különböző eseteket is vizsgálhatjuk. Így tehát ilyen kódokra is szükség van. Mivel minden nyelv rendelkezik sajátosságokkal, így lehetetlen lenne ugyanazt a rendszert „ráerőszakolni" az összesre.

A Brit Nemzeti Korpusz (BNC) elemzéséhez 61 alapkódot használtak http://www.natcorp.ox.ac.uk/what/ucrel.html. Az automatikus címkézés hibaszázaléka 1,7% körüli volt, és a szavak kb. 4,7%-át a program nem tudta egyértelműen azonosítani, így a korpusz 2%-át szakemberek utólag felülvizsgálták, és egy bővített, 160 kódból álló készlettel pontosították. Az eredmény kevesebb, mint 0,3% hibát tartalmaz. A BNC Sampler a kézi elemzésű alkorpuszt tartalmazza, melynek annotált szövege a következőképpen néz ki (written, world affairs, aa4 fájlból vett részlet):

```
<head type=MAIN>
<s n=0001 p=Y><w NP1>Jordan <w VVZ>lifts <w MC>22 <w
NNT2>years <w IO>of <w JJ>martial <w NN1>law<c YSTP>.
</s>
</head>
```

```
<head type=BYLINE>
<s n=0002 p=Y><w II>By <w NP1>Wafa <w NP1>Amr <w II>in
<w NP1>Amman </s>
</head>
<p>
<s n=0003 p=Y><w NP1>JORDAN<w GE>'S <w JJ>Prime <w
NN1>Minister<c YCOM>, <w NNB>Mr <w NP1>Mudar <w
NP1>Badran<c YCOM>, <w RT>yesterday <w VVD>announced <w
AT>the <w NN1>freezing <w IO>of <w JJ>martial <w
NN1>law <w IF>for <w AT>the <w MD>first <w NNT1>time <w
II>since <w MC>1967 <w II>as <w AT1>a <w N1>prelude <w
IF>for <w RR>formally <w VVG>revoking <w DAT>most <w
IO>of <w APPGE>its <w NN2>provisions <w CC>and <w
VVG>abolishing <w PPH1>it<c YSTP>. </s>
</p>
```

9. ábra: A BNC annotált szövege

Ebből is láthatjuk, hogy nem emberi szem számára készült, hiszen alaposan meg kell nézni, hogy a következő szöveget felfedezzük:

```
Jordan lifts 22 years of martial law.
By Wafa Amr in Amman
JORDAN'S Prime Minister, Mr Mudar Badran, yesterday
announced the freezing of martial law for the first time
since 1967 as a prelude for formally revoking most of
its provisions and abolishing it.
```

10. ábra: A BNC szövege annotálás nélkül

Egyrészt az olvashatóság kedvéért is praktikus mindkét változatban, annotálatlan és annotált változatban is megtartani az eredeti szöveget, másrészt meg kell említenünk, hogy a „kész korpuszok” felhasználói nem minden esetben értenek feltétlenül egyet az elemzés kategóriáival. Ha a kutató a későbbi kutatások során saját kódrendszerét kívánja használni ugyanannak a korpusznak az elemzésére, könnyebb azt egy annotációval nem rendelkező szövegbe illeszteni. Az azonos szövegre alkalmazott különböző kódrendszerek lehetővé teszik új szoftverek és új elemzési szempontok vizsgálatát és összehasonlítását is.

A grammatikai jellegű kódok mellett a szöveg eredetére vonatkozó információ is szerepel minden fájl fejlécében, és a szövegben mindvégig jelölik a szöveg megjelenésére (layout) és szerkezetére vonatkozó információt. Az írásjelek, a mondatok, bekezdések jelei is jelen vannak. A BNC annotált szövegét bemutató 9. ábra első sora a cikk címére vonatkozó információ kezdetét jelöli: <head type=MAIN>, majd a cím után a végét is jelöli </head>. A következő jelölés az új egység kezdetét jelöli, majd az információ után a végét jelölő sor következik. Így tehát a szövegre vonatkozó jelek mindig

közrefogják azt az egységet, amelyre vonatkoznak. A bekezdés, azaz paragrafus elejét jelzi, majd a végét , a mondatét <s> és </s>.

A Szerb Nyelv Korpuszának elemzése eredetileg kézzel és több mint 2000 kód felhasználásával készült, és még az 1950-es (!) években kezdődött. A kódokat 1998-ban felülvizsgálták, modernizálták és egyszerűsítették. Az eredeti kódolás a következőképpen néz ki http://www.serbian-corpus.edu.yu/ie/tagging/etagging.html:

A	B	C	D	E	F	G
Petar	**Petar**	i nom J mu	100111	5	2	10101
biti	**je**	gl prez-s-perf	523311	2	1	10
otići	**otišao**	gl rp-s-perf 3l J mu	522311	6	4	010100
u	**u**	pr	800000	1	1	0
škola	**školu**	i a J ž	100412	5	2	11010

Magyarázat: **Petar je otišao u školu.** – Peter iskolába ment.
A – maga a szó, **B** – eredeti szöveg, **C** – nyelvtani kód, **D** – nyelvtani kód numerikus formában, **E** – grafémák száma, **F** – szótagok száma, **G** – fonológiai szerkezet.
* **i** – főnév, **gl –** ige, **pr** – elöljárószó, **nom** – alanyeset, **a –** tárgyeset, **prez** – jelen idő, **perf** – múlt idő, **s** – valami részeként, **3l** – harmadik személy, **J –** egyes szám, **mu** – hímnemű, **ž –** nőnemű, **rp** – befejezett melléknévi igenév.

5. táblázat: A Szerb Nyelv Korpuszának elemzett formája

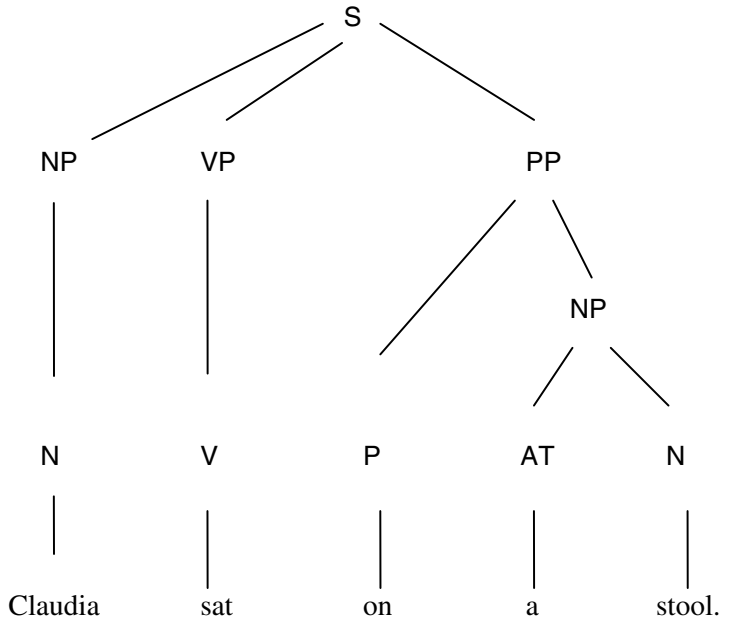
A mondattanilag elemzett korpuszok száma még kevesebb, mint a szófaji jelekkel ellátott korpuszoké. A mondattani elemzések két fajtája létezik: a „csontváz" elemzés (*skeleton parsing*) vagy „sekély" elemzés (*shallow parsing*) és a teljes elemzés (*full parsing*). Mint a megnevezésből sejthetjük, a teljes elemzés több információval fog szolgálni, mint a csontváz vagy sekély elemzés. Ez utóbbi nagyobb egységeket azonosít, és az egységen belül nem jelzi minden elem kapcsolatát. Nézzünk meg erre egy példát:

```
[ S [ N Nemo_NP1 ,_, [ N the_AT killer_NN1 whale_NN1 N ]
,_, [ Fr [ N who_PNQS N ][ V 'd_VHD grown_VVN [ J
too_RG big_JJ [ P for_IF [ N his_APP$ pool_NN1 [ P on_II
[ N Clacton_NP1 Pier_NNL1
N ] P] N ] P ] J ] V ] Fr ] N ] ,_, [ V
has_VHZ arrived_VVN safely_RR [ P at_II [ N  his_APP$ new_JJ
home_NN1 [ P in_II [ N Windsor_NP1 [ safari_NN1 park_NNL1
] N ] P ] N ] P ] V ] ._. S]
```

11. ábra: Példa a csontváz/sekély elemzésre a UCREL honlapjáról http://www.comp.lancs.ac.uk/computing/research/ucrel/annotation.html

Az eredeti mondat szavait és az elemzés egyes elemeit az olvashatóság kedvéért kövérrel szedtük. Az elemzés részletes leírása helyett csak bizonyos elemekre hívnánk fel a figyelmet, amely alapján mindenki kedvére „megfejtheti" a többit. Az első zájójelet követő **S** és az utolsó zárójel előtti **S** fogja közre a teljes mondatot. Az első sor vége felé található **Fr** és a harmadik sor közepén található párja jelölik a vonatkozó mellékmon-

datot. Tehát a jelek és a zárójelek közrefogják azokat az elemeket, amelyek azt az egységet alkotják. Az elemek több egység részét képezik, így a zárójelek többszörösen közrefogják őket. Ahhoz, hogy a teljes elemzés pontosabb képet adjon, még több azonosító jelre van szükség. Ezeket az elemzéseket a zárójelek helyett a generatív nyelvészeti elemzésekből jól ismert ágrajz, amit angolul *tree*-nek neveznek, segítségével is ábrázolhatjuk, ezért az ilyen szintaktikai elemzéseket tartalmazó korpusz neve *treebank.*



12. ábra: Példa az ágrajzra

A mondattani elemzés eredményeiről és nehézségeiről számos szakcikket olvashatunk. Minden nyelvnek megvannak az elemzési nehézségei, így a publikációk nem általános, hanem egyes nyelvekhez kapcsolodó gondokról számolnak be. Váradi Tamás (2003) a magyar üzleti szövegek mondattani elemzésének egyes problémáiról ír. Ahhoz, hogy magyar nyelvű „treebank"-et létrehozhassanak, meg kell teremteni a megfelelő számítógépes programcsomagokat. Jelenleg is folynak a kutatások ezen a területen.

Az előzőekben a nyelvészeti szempontból standard, azaz általánosan elfogadott, megszokott annotációkról volt szó. Mivel az annotáció és az elemzések számítógéppel történnek, így a technikai standardokról is említést kell tennünk. Könnyen belátható, hogy minden korpuszfelhasználó érdeke azt kívánja, hogy a meglévő számítógépes programok mások által készített annotált korpuszok elemzésére is képesek legyenek. Ezért a *szövegkódolási ajánlás* (Text Encoding Initiative, TEI)[14] mindenkinek a *szabvá-*

[14] A TEI olyan nemzetközi és interdiszciplináris szabvány, amely a szövekódolást igyekszik egységessé tenni. A pontosságra és egyszerűségre törekednek. A szabvány fejlesztésére 1987-ben konzorciumot hoztak létre, amely 2000-ben megújult (http://www.tei-c.org/).

nyos általános leíró nyelv (Standard Generalised Markup Language, SGML)[15] használatát ajánlja. Számos publikáció nyújt részletes információt az SGML szövegkódolásban való használatáról (Lou Burnard, 1992, 1995; Sperberg-McQueen, 1994; Sperberg-McQueen & Burnard, 1994).

1.3.4.3. Speciális annotációk

A szófaji és mondattani annotációk a legelterjedtebbek, de ezeken kívül mások is léteznek, és valószínű, hogy számuk egyre nő aszerint, hogy milyen elemzéseket kíván a korpusz alkotója elvégezni a szövegen. Léteznek ortografikus, fonetikus/fonémikus, prozódikus (8. ábra), szemantikai (13. ábra), diskurzus (14. ábra), pragmatikus/stilisztikai annotációk (Leech & Eyes, 1997: 12). Bárki – saját szükségletének és ötletességének megfelelően – bármilyen annotációt alkalmazhat a korpusz bizonyos szempont vagy szempontok szerinti elemzésére. Csak azt az elvet kell szem előtt tartani, hogy egyértelmű legyen mind a jelölési rendszer, mind pedig az, hogy a címke melyik elemre vonatkozik.

There_Z5's_Z5_been_A3+ more_NS++ violence_E3- in_Z5 the_Z5 Basque_Z2 country_M7 in_Z5 northern_M6 Spain_Z2 :_PUNC one_N1 policeman_G2.1/S2m has_Z5 been_Z5 killed_Ll- ,_PUNC and_Z5 two_N1 have_Z5 been_Z5 injured_B2- in_Z5 a_Z5 grenade_G3 and_Z5 machine-gun_G3 attack_G3 on_Z5 their_Z8 patrol-car_M3/G2.1 ._PUNC

13. ábra: Szemantikai annotáció (Leech 1997: 13)

(0) The state Supreme Court has refused to release {1 [2 Rahway State Prison 2] inmate 1}} (1 James Scott 1) on bail .

(1 The fighter 1) is serving 30-40 years for a 1975 armed robbery conviction . (1 Scott 1) had asked for freedom while <1 he waits for an appeal decision. Meanwhile, [3 <1 his promoter 3] , {{3 Murad Muhammed 3}, said Wednesday <3 he netted only $15,250 for (4 [1 Scott 1] 's nationally televised light heavyweight fight against {5 ranking contender 5}) (5 Yaqui Lopez 5) last Saturday 4) .

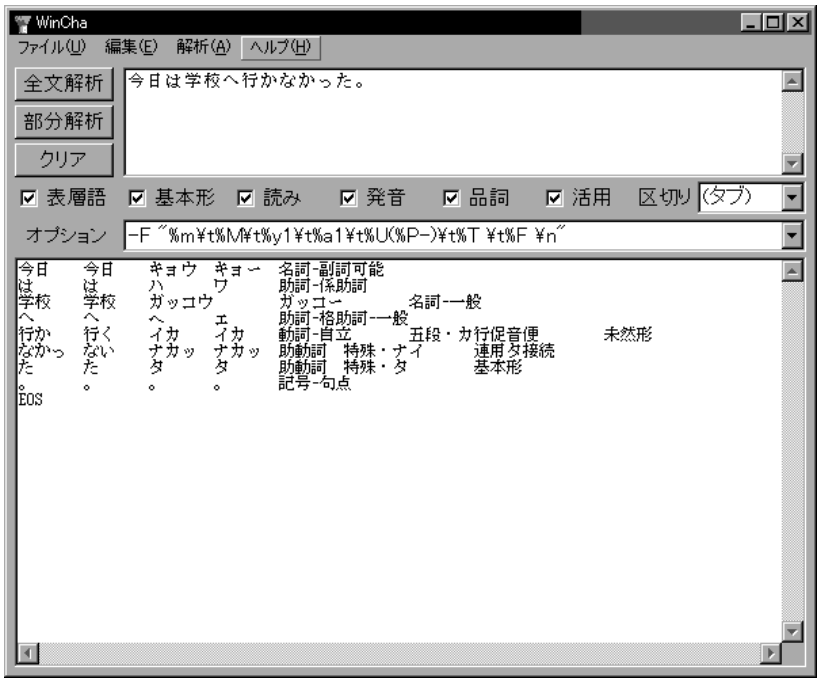
14. ábra: Diskurzus annotáció (Leech 1997: 13)

A korpusz módszerek terjedésének köszönhetően egyre több pedagógus állít össze tanulói korpuszt, hogy világosabban lássa, milyen hibákat követnek el diákjai a nyelvtanulás során. Természetesen az ilyen „hibafeltáró” korpusz esetében az annotációnak a hibákra vonatkozó információt kell tartalmaznia. Az ilyen jellegű korpuszok közül talán Tono Yukio japán nyelvész és nyelvtanár által összeállított 700 000 szóból álló korpusz a legismertebb. 640 diák hat megadott témáról szabadon írt angol nyelvű fogalmazását

[15] Fordítása: szabványos, kiterjesztett jelölő nyelv. Szöveges állományok belső szerkezetének (fejezetek, bekezdések, lábjegyzetek stb.) jelölésére használható szabvány. A HTML nyelv például eredetileg az SGML egyik egyszerűsített változata volt (http://www.bibl.u-szeged.hu/mke_eksz/docs/ekszotar/s.htm).

tartalmazza. Szinte mindenki automatikusan idegen nyelvekre gondol, amikor hibákról van szó, pedig anyanyelvünk tanulása során is rengeteg hibát vétettünk. Így a hibafeltáró korpusz az anyanyelv eredményesebb tanításában is nagy szerepet játszhatna.

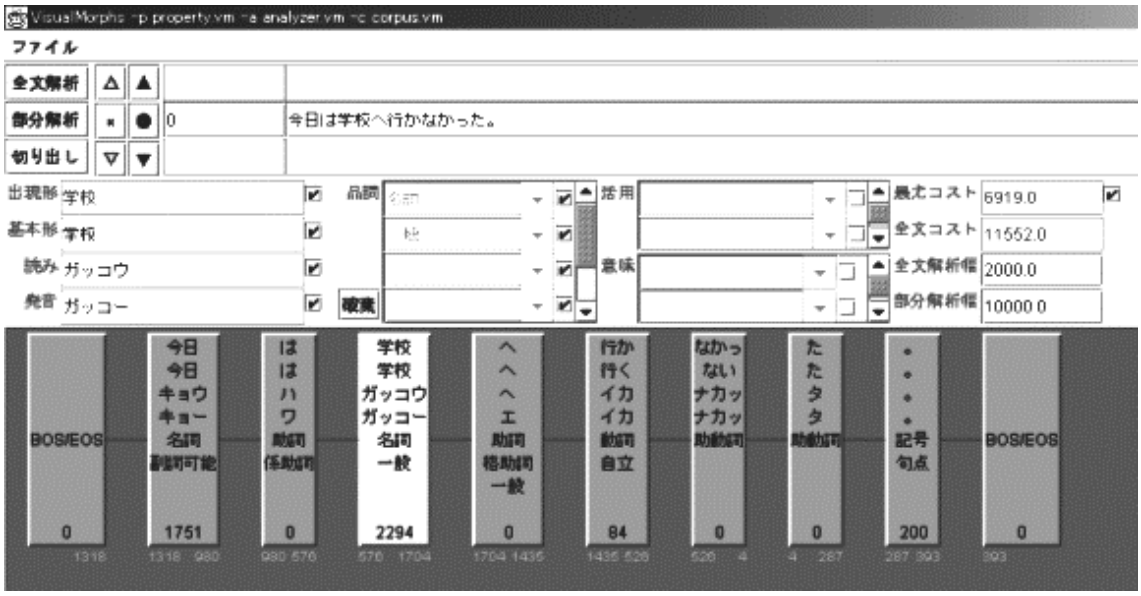
Mivel a korpusznyelvészet kialakulása során az angol nyelv elemzése volt szinte mindvégig a figyelem központjában, a „szabványos" annotációk is leginkább csak az angol nyelv elemzését tartják szem előtt (Oravecz & Dienes, 2002). Ezzel szemben minden egyes nyelvnek vannak olyan problémái, amelyeket lehetetlen az ajánlott kódrendszerrel leírni. Ilyen esetben saját annotációra vagy a standard kiegészítésére, módosítására van szükség. Viszonylag könnyű az izoláló nyelveket annotálni, amelyek esetében a grammatikai információt leggyakrabban külön szó fejezi ki (pl. az angol), és nem pedig a szavak belsejébe ékelődő morfémák (pl. magyar, japán). Az agglutináló nyelvek esetében morfológiai annotációra is szükség van. Például a lehetőség kifejezésére a magyar nyelvben a *-hat/-het* képzőt használjuk (pl. olvas*hat*, e*het*), a japánban pedig a *-(r)e, -rare* (pl. yom*e*masu tabe*rare*masu) morfémákat. Vannak ugyan morfológiai elemző programok, de ezek eredményét is annotációként kell megjeleníteni a korpuszban minden egyes szóra vonatkozóan. Két ilyen morfológiai elemző programot említünk meg: a ChaSen nevű programot a Nara Institute of Science and Technology (Matsumoto et al., 1999) fejlesztette ki a japán nyelv elemzésére, a Prószéky és kollégái által készített HuMor pedig (Prószéky & Tihanyi, 1992, 1993) a magyar nyelvhez készült, és a számítógépes helyesírási elemző program részeként működik, de önálló programként nem forgalmazzák. A ChaSen azonban önálló programként letölthető, így az érdekesség kedvéért nézzünk meg egy példát ennek segítségével. A japán nyelv esetében még azzal a problémával is meg kell birkózni, hogy a szavak között nincs szóköz. A szó megszokott meghatározása ugyanis úgy szól, hogy két szóköz által határolt egység.



15. ábra: Egy japán mondat morfológiai elemzése

A képernyő felső részén látható az eredeti mondat, melynek fordítása így hangzik: *Ma nem mentem iskolába.* Latin betűs átirata pedig a következő: kyouwagakkoueikanakatta. A kép alsó részében függőleges oszlopok láthatók, melyek az egybefolyónak látszó szöveget alkotóelemeire bontják. A bal oldali első oszlopban szerepel tehát a mondat morfológiai elemekre bontva, mely így hét elemre tagozódik. A második oszlopban az egyes elemek alapformája, a harmadik oszlopban az olvasata, a negyedikben a kiejtése, az ötödikben a szófaja, legvégül pedig a használatára vonatkozó információ következik.

Ugyanennek a mondatnak a VisualMorphs (Matsuda *et al.*, 2001) nevű programmal történő elemzése a következő képet eredményezi:



16. ábra: Morfológiai elemzés vizuális elrendezésben

Ez a morfológiai elemzés is ugyanazt az eredményt mutatja, mint az előző ábra, de az információ megjelenítésének módja sokkal könnyebbé teszi ennek áttekintését. A 15. ábra estében nem egyértelmű, hogy az összetartozó adatok vízszintesen vagy függőlegesen olvasandók-e. A 16. ábra viszont egyértelmű a japánul nem tudó olvasó számára is. Az egyes elemekre vonatkozó információk jól áttekinthetők. A „szavak” az eredeti mondatnak megfelelő sorrendben és vízszintes irányban olvashatók. Ha az egér mutatóját az elemzést megjelenítő részben egy elem fölé mozdítjuk, az arra az elemre vonatkozó információk a kép közepén látható ablakocskákban jelennek meg.

1.4. Összefoglalás

Ebben a fejezetben a *korpusz* számos meghatározását olvashattuk, melyekben a közös tulajdonságok a következők voltak:

1. elektromos formában tárolt szövegek gyűjteménye;

2. a szövegeket meghatározott szempontok alapján válogatják és reprezentatívnak, azaz jellemzőnek találják;
3. legtöbbször bizonyos szempontok szerint annotációval látják el, hogy a lekérdezést gyorsabbá és pontosabbá tegyék;
4. a korpusz nyelvészeti elemzés céljából készül, így az eredménye is nyelvészetileg értékes információkkal szolgál.

A reprezentativitást a nyelvészeti célok tükrében lehet csak értelmezni. Ha valaki a mai magyar diáknyelvet kívánja vizsgálni, nem használhat vizsgálódásaihoz egy tankönyvek szövegéből álló korpuszt, hiszen az nem reprezentatív arra nézve, hogy a diákok valójában hogyan is beszélnek. Ezen kívül a korpusz méretének is döntő szerepe van. Általánosan elfogadott nézet az, hogy a nyelvtan tanulmányozásához kisebb korpusz is elegendő, akár egy egymillió szóból álló korpusz, de az általános nyelv lexikográfiai vizsgálatához a lehető legnagyobb korpuszra van szükség. A „minél nagyobb, annál jobb” elvet kell itt követni.

A vizsgálat céljától függően a mintavétel módjára is ügyelni kell. Számos korpusz esetében csak bizonyos mennyiségű szó kerül egy szövegből a korpuszba, nem pedig a teljes szöveg. Az ilyenfajta mintavétel nem elfogadható olyan esetekben, amikor a szöveg egészére vonatkozó megállapításokat kívánunk tenni.

A korpuszokat a mintavétel módja és a felhasználás alapján csoportosítottuk. Így beszéltünk statikus, dinamikus és monitor korpuszról, valamint általánosról és speciálisról. Meghatároztuk és példát adtunk az összehasonlítható, párhuzamos, fordítási, nyelvtanulói, pedagógiai és történelmi vagy diakrón korpuszra. A jogi problémák megemlítése után a korpuszokban szereplő standard és speciális annotációval fejeztük be a korpusz fogalmának és természetének ismertetését. Végül pedig példát mutattunk be a morfológiai elemző programra.

2. SZÁMÍTÁSTECHNIKA ÉS NYELVTUDOMÁNY

2.1. Bevezetés

Az első fejezetben a korpusznyelvészet és a korpusz meghatározása után a korpusz általános jellemzőiről szóltunk. Az eddigiekből világosan kitűnik, hogy a korpusznyelvészet számítógépek nélkül nem létezhet, így ezt a fejezetet a technikai fejlődés rövid ismertetésével kezdjük. A korpuszok és a korpusznyelvészet fejlődése a technikai háttér minimális ismerete nélkül nem igazán értékelhető. Ezek után a szellemi, azaz nyelvészeti hátteret mutatjuk be, majd kapcsolódó tudományágakról és a számítógépes nyelvészet hazai fejlődéséről esik röviden szó.

2.2. A számítástechnika fejlődése

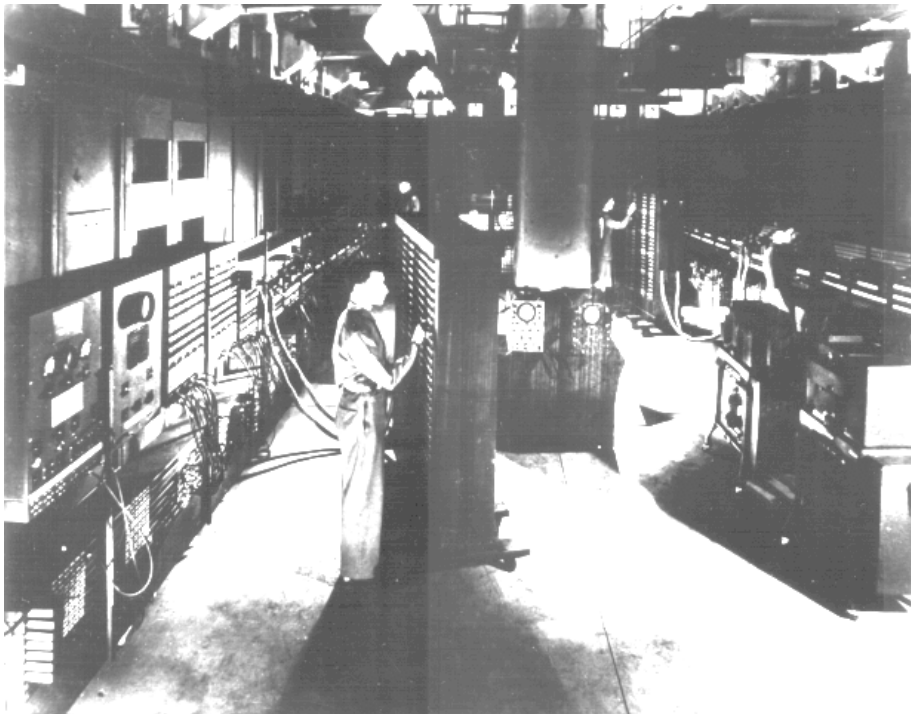
Jóllehet már az 1960-as évek előtt is készült nyelvészeti elemzés céljára néhány korpusz, ezeket azonban nem elektronikus formában tárolták, így az adatok elemzése és feldolgozása nem történhetett számítógépes programok segítségével, hanem csak kézzel és papíron végezték. Nem is lehet ezen csodálkozni, hiszen a számítógépek története sem nyúlik időben sokkal hátrább, mindössze az 1930-as, 40-es évekig. Mivel a korai korpusznyelvészet fejlődését meghatározta és behatárolta a számítógépek hardverének és szoftverének limitált teljesítőképessége, szükségesnek tűnik, hogy – ha nem is a teljesség igényével – egy pár szót szóljunk a számítógépek fejlődéséről. A korpusznyelvészet fejlődését és az úttörő munka nehézségeit is csak így lehet értékelni és megérteni.

Még számítástechnikai szakemberek számára is nehéz feladat eldönteni, hogy igazából mit is illethetünk „az első számítógép” névvel. A technika és tudomány történetében többször előfordult, hogy nagyjából azonos időpontban, de egymástól függetlenül két feltaláló kísérleteit is siker koronázta. A számítógép létrehozásához szükséges elméleti és gyakorlati ötletek közül számos Pascalig is visszavezethető. Ha a szabadalmak alapján szeretnénk ezt a vitát eldönteni, akkor sem járnánk sok sikerrel, mert vagy két évtizeden keresztül folytak perek e téren. Elégedjünk itt meg a köztudatba bekerült és elfogadott eseményekkel. A kíváncsiabbak figyelmébe ajánlom McCartney könyvét az ENIAC-ról (1999) és Goldstine (1993) könyvét a számítógép történetéről.

A legtöbb korai próbálkozással az volt a baj, hogy a gép emberi közreműködés nélkül csak előre meghatározott műveleteket tudott végrehajtani, sokszor egyszerre csak egyfélét, és a kapott eredményeket nem tudta újabb műveletre felhasználni, hanem azokat újra be kellett táplálni a gépbe. Az első elektronikus számítógép, az ABC (Atanasoff-Berry Computer), 1942-ben jelent meg, amit az ENIAC követett 1944–45-

ben. Az ENIAC sokkal többet „tudott” bármely más gépnél, hiszen másodpercenként 5000 összeadást vagy 300 szorzást volt képes elvégezni, míg más gépek csak egy szorzási műveletre voltak képesek ugyanennyi idő alatt. Manapság ez a szám valószínűleg több százmillió, attól függően, hogy milyen sebességű a központi processzor egység, azaz CPU (Jamsa, 1997). E mellett az ENIAC képes volt 20 darab 10 karakterből álló szó „gyors” memóriában és 450 szó ROM-ban való tárolására. Ennek ellenére egyik problémáról egy másikra való átállás így is órákat vagy napokat vett igénybe. Ezek után, a von Neumann, azaz Neumann János és H. H. Goldstine által az Institute for Advanced Study (IAS)-nál (Princeton, New Jersey) vezetett projekt eredményeképpen 1952-ben olyan gépet készítettek, amely a modern elektronikus számítógép prototípusa lett.

A kereskedelmi forgalomban vásárolható számítógépeket sokszor emlegetik „generációkra” utalva az alapján, hogy milyen technológiát használtak gyártásukhoz. Az első generációs számítógépek (1952–1958), elektroncsöves technológiával (vacuum-tube) készültek, elsősorban nagy, természettudományi tevékenységet folytató vevők, mint például a kormány által támogatott laboratóriumok számára. A legelső sorozatgyártásban készült gépek a UNIVAC mellett az IBM 701 és 702 voltak.



17. ábra: Az ENIAC 2

A második generációt tranzisztoros technológia jellemezte, 1959 és 1963 között került forgalomba. A harmadik generáció (1964-től 1975-ig) időszakában jelentek meg az integrált áramkörök. Az első számítógépes „család” is csak 1964-ben jelent meg (IBM

System/360), mely azt jelentette, hogy ugyanazt a programot a család minden egyes tagján futtatni lehetett. 1972-ben találták fel a csipeket, amelyek több integrált áramkör kombinációjából állnak.

A negyedik generációs számítógépekre (az 1970-es évek közepétől az 1980-as évek közepéig) a nagyon nagy méretű áramkörök, multiprocesszorok és hálózatok voltak jellemzők. A memóriacsipek olcsóbbak lettek és megjelentek a nagy felbontású képernyők, valamint a kép orientált programok. Az új áramkörök lehetővé tették a miniszámítógépek, a nagy teljesítményű személyi számítógépek és a munkaállomások kifejlesztését is. Ezzel a személyi számítógépek a nagyközönség számára is elérhetővé és megfizethetővé váltak.

Az Altair 8800, az első széles körben használt számítógép 1974-ben került piacra, majd ezt követte a Commodore PET 1977-ben és a Radio Shack által gyártott Tandy. Az Apple Computer céget 1976-ban alapították és 1980-ra az éves eladásuk már 100 millió dollár felett volt. Az IBM cég Intel 8086-os processzorral rendelkező személyi számítógépe 1981-ben jelent meg, és nem sokkal ezután a cég engedélyezte, hogy más cégek is gyártsák ennek olcsó „klónjait", valamint arra ösztönözték a szoftverfejlesztő mérnököket, hogy írjanak programokat az IBM személyi számítógépére.

Az ötödik generáció (1982–1994) figyelmét a nem numerikus adatfeldolgozásra, alkalmazott mesterséges intelligenciára, és az intelligens interfészekre irányította. Jóllehet a vizsgált problémákat nem sikerült ezen időszakban megoldani, számos technikai előrelépés történt a természetes nyelvek feldolgozása, a képfeldolgozás és a hálózatok terén. Az 1990-es évek elejétől kezdve a technikai fejlődés jelentősen felgyorsult. Szinte félévente egyre gyorsabb és gyorsabb processzorok kerülnek forgalomba. Már a notebook és a kézben tartható (handheld) számítógépek is gyorsabbak és nagyobb adattárolási kapacitással rendelkeznek, mint a korai korpuszelemzések idején használt hálózati (mainframe) számítógépek. Az Encyclopedia Americana Version 5.0 („Encyclopedia Americana", 1999) által közölt adatok (6. táblázat 2. és 3. oszlopa) az első kereskedelmi forgalomban beszerezhető UNIVAC I technikai paramétereit hasonlítja össze az első személyi számítógép paramétereivel. A technikai adatok mellett az ár is fel van tüntetve, mert ebből következtetni lehet arra, hogy kik lehetnek a potenciális vásárlók, és mennyire elterjedt vagy ritka lehet az adott technika. A laptopokra vonatkozó adatok tájékoztató jelleggel szerepelnek, hiszen a rengeteg különböző gyártmány közül nehéz lenne egyet kiválasztani. Az említett adatok közül manapság már nem használják az utasítás/másodpercet, az árak is igen változatosak, és a gép fogyasztása nagymértékben függ a használt eszközöktől (nyomtató, CD-író stb).

	UNIVAC I	IBM PC	Laptop
Megjelenés éve	1951	1981	2004
Sebesség (utasítás/másodperc)	20 000	400 000	Több 100 millió
Ár (dollárban)	250 000	5000	1000–2000
Energiafogyasztás (watt)	50 000	500	75–150

6. táblázat: A kereskedelmi forgalomban levő elektronikus számítógépek első 30 éve: technikai összehasonlítások

Ez a gyors fejlődés az utóbbi 5-10 évben még jobban érezhető. Hadd említsek itt egy személyes példát. Az első általam használt asztali számítógép, 256K RAM memóriával rendelkezett 1992-ben. Az első notebook számítógépem 1994-ben 120 Mb-os merevlemezzel rendelkezett. Manapság a kulcstartóra fűzhető flash memóriák, vagy pendriveok is sokkal több adatot tárolnak. Az 1996-ban vásárolt Toshiba számítógépem, amelyre oly büszke voltam akkor, ma már nemcsak hogy eladhatatlan, de még ajándékba sem nagyon akarná senki sem elfogadni. A jelenlegi – már háromévesen öregecske – notebook számítógépem most 256Mb RAM-mel, 1,19 GHz-es processzorral és 40Gb merevlemezzel rendelkezik. Változnak az idők!

A korpusznyelvészet szempontjából azonban nem csak az adattárolás és az adatfeldolgozás sebességének fejlődése volt meghatározó, hanem az adatbevitelé is. A korai korpuszok készítésekor jelentős akadályt jelentett az adatbevitel nehézsége. John Sinclair korai, Edinburghban, Manchesterben és Birminghamben végzett munkája során (J. Sinclair *et al.*, 1969) a beszélt nyelvi adatokat átírták, majd kézzel lyukkártyára vitték az adatokat és egy kártyaolvasó segítségével olvasták be a számítógépbe. Ez volt az egyik oka annak, hogy ez a korpusz mindössze 300 000 szóból állt (J. M. Sinclair, 2001). Az 1970-es években még mindig használtak lyukkártyákat a programok bevitelére, de az adatokat már kicsit gyorsabban, mágneses szalagról vagy papír lyukszalagról olvasták be. A birminghami Cobuild Korpusz esetében még 1980–83 között is gépelni kellett az adatokat, így 1984-ben is mindössze 7,3 millió szóból állt.

1984-ben a Birminghami Egyetem egy Kurzweill Data Entry Machine-t (KDEM), azaz optikai szkennert szerzett be, amellyel az adatbevitel sokkal gyorsabbá vált, és 1985-re a korpusz már kb. 18–20 millió szóra növekedett. Az 1980-as évek vége felé és az 1990-es évek elején egyre több anyagot kaptak mágneses szalagon, így például a Times és a BBC anyagát. Napjainkban a kézi adatbevitel, illetve a szkennerek használata háttérbe szorult, hiszen az adathordozók óriási kapacitása és az internetet gyorsító technológiák lehetővé teszik, hogy akár egyszerű fájl-átviteli eljárással (FTP) bővítsük az adatbázisokat. Az internetről tehát még az egyszerű érdeklődő is rengeteg adathoz juthat egy korpusz létrehozásához.

2.3. A korpuszok fejlődése

Az első kereskedelemben kapható számítógép 1951-ben jelent meg, és az első számítógépes korpusz munkálatait 10 évvel később kezdték meg. Szintén ekkorra készült el a világ leggyorsabb számítógépe, a Stretch, mely magas ára miatt nem válhatott sikeressé. A második generáció ideje ez, amikor az amerikai légitársaság, az American Airlines, az IBM-mel együttműködve létrehozza a Sabre nevű utazási helyfoglaló rendszert 1962-ben. Az Assembly programozási nyelv helyett magasabb szintű programnyelvek jelennek meg, mint például a FORTRAN vagy a COBOL (Nadar, 1998). Ebben az időben, és ilyen technikai feltételek között született meg az első elektronikus korpusz, a Brown University Standard Corpus of Present-Day American English, azaz a Brown Egyetem Mai Amerikai Angol Nyelvének Standard Korpusza, melyet mindenki csak Brown Korpuszként emleget.

2.3.1. A szellemi háttér

A technikai háttér lehetőséget ad a gondolatok megvalósítására, de mint oly sok más területen, a nyelvészetben is vannak kedvező és ellenséges szellemi erők és divatok, melyek erősen befolyásolják az új gondolatok fogadtatását. Noam Chomsky *Syntactic Structures* (1957) című műve megjelenésének eredményeképpen, különösen az Egyesült Államokban, a nyelvészek többsége évtizedeken keresztül a transzformációs-generatív nyelvészettel foglalkozott. Ezen irányzat teljesen figyelmen kívül hagyja a tényleges nyelvhasználat vizsgálatát, és céljának elsősorban a nyelv belső szabályrendszerének leírását tartja. Mivel a generativista irányzat a beszélő megnyilatkozását egy tökéletes nyelvtan tökéletlen, hibákkal tarkított megnyilvánulásának tekintette, tényleges megnyilvánulásokat, szövegeket nem vizsgált, és így az empirikus módszereket is megvetette. Az ilyenfajta vizsgálódást merő időpocsékolásnak, vagy az állami pénzek elherdálásának tekintették. Úgy tűnik, hogy Chomsky véleménye egyáltalán nem változott e tekintetben a majdnem 50 év alatt sem, hiszen egy Bas Aartsnak adott interjúban (2000) még a korpusznyelvészet puszta létét is tagadta. Az *Aspects of the Theory of Syntax* című művében ezt írja: *„A nyelv ismerete, mint a legtöbb fontos és érdekes tény, se nem figyelhető meg közvetlenül, se nem vonható ki adatokból semmilyen induktív módszer segítségével”*[16] (1965: 18). Andor József személyesen készített interjút Chomskyval 2004 januárjában és ekkor a korpusznyelvészetre vonatkozó kérdésre Chomsky így kezdte válaszát: *„A korpusznyelvészetnek nincs értelme”*[17] (2004: 97).

Egy ismert amerikai nyelvész, Charles Fillmore a generatív és korpusznyelvészt kicsit parodikusan a következőképpen írta le: A generatív vagy elméleti nyelvész egy kényelmes fotelban ülve, csukott szemmel elmélkedik, majd szemét kinyitva felkiált: Milyen érdekes tény! Felkapja a tollát, és jegyzetelni kezd. Öröme órák múltán sem csillapul, hiszen egy lépéssel közelebb került a nyelv megismeréséhez. A korpusznyelvész viszont hatalmas korpuszában valójában minden elsődleges tény birtokában van. Céljának azt tekinti, hogy az elsődleges tényekből másodlagos tényeket hozzon létre (1992: 35). Érdemes itt ismertetnünk egy másik összehasonlítást is, melyet Lager (1995: 3) doktori dolgozatából vettünk:

A KORPUSZNYELVÉSZ	AZ ELMÉLETI NYELVÉSZ
Figyelmét a **beszéd/performancia/folyamat**nak szenteli.	Figyelmét a **nyelv/kompetencia/rendszer**nek szenteli.
Célja a nyelv tényeinek és a nyelvhasználatnak a korpuszban való megjelenés szerinti leírása.	Célja az egyén tudatában élő nyelv tényeinek a magyarázatát adni.
A nyelvészeti kutatásai **adat-vezéreltek**. Imádja a hosszú listákat.	Nyelvészeti kutatásait **elméletek** vezérlik.
Megelégszik egyszerű módszerekkel, mert gyorsak, még hosszú szövegek esetében is. Úgy érzi, hogy a módszerei egyszerűsége által teremtett pontatlanságot el lehet viselni.	Az egyszerűségnél jobban kedveli a kifinomultságot, a sebességnél a pontosságot, még ha ez azzal is jár, hogy kevesebb példával is be kell érnie.

[16] Eredetiben: *“Knowledge of the language, like most facts of interest and importance, is neither presented for direct observation nor extractable from data by inductive procedures of any known sort.”*
[17] Eredetiben: *“Corpus linguistics doesn’t mean anything.”*

A KORPUSZNYELVÉSZ	AZ ELMÉLETI NYELVÉSZ
Kedveli a **kvantitatív** módszereket.	Kedveli a **kvalitatív** módszereket.
Önmagát az **empirikus** hagyományok követőjének tekinti.	Önmagát a **racionalista** hagyományok követőjének tekinti.
A szöveget **fizikai produktumként** kezeli.	A szöveget **absztrakt entitásként** kezeli.
Főként egyes nyelvek nyelvtana érdekli.	Az egyetemes nyelvtan érdekli.
Csak a **formára** figyel.	A formára **és** a jelentésre figyel.
A szövegeket **globálisan** vizsgálja. Át akarja tekinteni, hogy mi van a szövegben.	A szövegeket **lokálisan** vizsgálja, azaz a részletekre összpontosít.
Megkötésektől mentes szövegek széleskörű (bár valószínűleg felszínes) elemzésére összpontosít.	(Mesterségesen) megkötött területek mély elemzését végzi.
Kedveli a statisztikai és a valószínűség elméletén alapuló módszereket.	Kedveli a **szabályvezérelt** (tudásalapú, dedukciós, logikán alapuló) módszereket.
A nyílt nyelvi viselkedés (produktumainak) megfigyelésére támaszkodik.	Fő adatszerző módszere az intuícióra/önvizsgálatra támaszkodik.
Autentikus adatokkal és más beszédmegnyilvánulások kontextusában levő beszédmegnyilvánulásokkal dolgozik.	„Játék példákkal" dolgozik izolált, kontextus nélküli mondatok formájában.
Bacon gondonkodásmódját követi. Úgy gondolja, hogy a tudomány lényege az induktív módszer.	Úgy véli, hogy a feltételező-dedukciós módszer a tudomány lényege.
Szilárdan hisz a FELFEDEZŐ folyamatokban.	Az IGAZOLÓ/ÉRTÉKELŐ folyamatokban hisz.
Ha a számítógépet a szövegszerkesztésen kívül másra is használja, akkor azt konkordanciák készítése, lemmatizálás vagy statisztikai számlálás céljából teszi.	Ha érdekli egyáltalán a számítógép, akkor parserekkel, generátorokkal, és teoréma bizonyító programokkal dolgozik.
Ha egyáltalán foglalkozik programozással, akkor azt Pascal, C vagy Pearl nyelvben teszi.	A Prolog vagy Lisp az, ami számít, ha programozással is foglalkozik.

7. táblázat: A korpusz és az elméleti nyelvész paródiája (Lager 1995: 3)

Mint minden karikatúra, ez a leírás is alapjaiban igaz, de túlzó. A kétfajta nyelvészeti szemlélet ellentétének alapja tulajdonképpen a beszéd/performancia/folyamat és a nyelv/kompetencia/rendszerről vallott nézeteken alapszik, mely egyben a további különbségek kiváltó oka is. A performancia és kompetencia részletes vizsgálatát számos cikk taglalja (Fromkin, 1968; Milroy, 1985; Newmeyer, 1990; Taylor, 1988). Jóllehet ezt a fajta kettősséget a főbb modern nyelvészeti irányzatok elfogadták (Stubbs, 1996), elsősorban brit nyelvészek, mint Firth (1957), Halliday (1978) és Sinclair (1991) elvetették. A korpusznyelvészek a brit nyelvészeket követték.

A közmondás szerint is az „arany középutat" kell követni, melyet Fillmore úgy fogalmazott meg, hogy „két nyelvész egy testben" (1992: 35) kellene, hogy létezzen. Tehát, az elméleti nyelvésznek is szüksége van autentikus példákra és adatokra, míg a korpusznyelvész sem elégedhet meg puszta statisztikákkal, és az intuíciójára is támaszkodnia kell. Számos nyelvész fogadta meg Fillmore tanácsát mindkét táborból. Az elméleti nyelvészeti kutatási programok, mint az MIT Lexikon Projektje, a FrameNet, és a WordNet is használtak korpuszokat. A kutatók közül sokan kutatásaikhoz empirikus adatokat használtak, például Levin & Rappaport Hovav (1995), Miller & Fellbaum (1992), Jackendoff (1977), Pinker (1989, 2000), és Pustejovsky (1995). Az empirikus

adatok és korpuszok használata elsősorban a szemantika, különösen a lexikális szemantika területén vált elfogadottá. A szövegelemzés és textológia is sokat profitált a korpuszok használatából. Ezzel egyidőben a korpusznyelvészet is sokat változott. Véget vetettek a „szószámlálásnak", és egyre kvalitatívabb lett. A kétfajta szemlélet inkább kiegészíti egymást, semmint ellentmondana egymásnak.

2.3.2. A korpusznyelvészet és a kapcsolódó tudományágak

2.3.2.1. A számítógépes nyelvészet

A korpusznyelvészet a legszorosabb kapcsolatban a számítógépes nyelvészettel van, és sokszor igen nehéz a határvonalat meghúzni köztük. A számítógépes nyelvészetet az Encyclopedia Britannica Online úgy határozza meg, hogy az pusztán az „elektromos digitális számítógépek használata a nyelvészeti kutatásban"[18] (Linguistics, 2005). Két szempontból tűnik ez a meghatározás túl egyszerűnek. Először is azt az érzetet kelti, hogy ez semmiben nem különbözik a hagyományos nyelvészettől, pusztán a számítógép használatában. Ha ez igaz lenne, akkor más tudományok esetében is, ha nem is minden esetben, ezt külön jelzővel illetnék, és beszélhetnénk számítógépes fizikáról vagy számítógépes kémiáról is, hiszen ott is használnak számítógépeket a kutatásban. A meghatározás alapján tehát a hagyományos értelemben vett nyelvészet „elnyeli" a számítógépes nyelvészetet, és ezek alapján nem lehet önálló diszciplína. Ez egyben magában hordozza a másik problémát. Azt az érzetet is kelti, hogy a számítógépes nyelvészet kutatási problémái azonosak a hagyományos nyelvészetével, csak számítógépet használnak vizsgálatukhoz. Ezzel szemben a számítógépes nyelvészet kérdései mások, mint a hagyományos nyelvészeté, és ezek vizsgálalát nem lehetne a számítógép segítsége nélkül elvégezni, tehát a számítógép használata szerves része, feltétele a kutatásnak.

Az emberek igen kis százaléka rendelkezik nyelvészeti szakszótárral, így kíváncsiságukat legtöbbször enciklopédiák segítsével elégítik ki, ami tévedésekhez, pontatlanságokhoz vezethet. Lássuk tehát a számítógépes nyelvészetnek egy nyelvészeti szótárban megjelent meghatározását is: *„A nyelvészetet és az (alkalmazott) számítógépes tudományokat egyesítő tudományág, mely a természetes nyelvek számítógépes feldolgozásával foglalkozik (a nyelvi leírás minden szintjén)"*[19] (Bussmann, 1996: 91).

2.3.2.2. A mesterséges intelligencia

A számítástechnika területéről a számítógépes nyelvészek számára elsősorban a mesterséges intelligencia kutatásai fontosak, melyeknek célja „az emberi intelligencia és kognitív képességek megértése és szimulálása számítógépek segítségével"[20] (ibid. 38). Jóllehet a mesterséges intelligenciával foglalkozó kutatások már a számítógépek meg-

[18] Eredetiben: *"no more than the use of electronic digital computers in linguistic research"*.
[19] Eredetiben: *"Discipline straddling linguistics and (applied) computer science that is concerned with the computer processing of natural languages (on all levels of lingusitis description)."*
[20] Eredetiben: *"to simulate and understand human intelligence and cognitive abilities by using machines (i.e. computers)"*.

jelenése előtt megkezdődtek, igazából a számítógépek megjelenése után nőtt meg az ezzel foglalkozó kutatók száma. Az első mesterséges intelligenciáról szóló tudományos cikket 1950-ben Alan Turing angol tudós publikálta. Az első főállású alkalmazottakból álló kutatócsoport 1954-ben, a Carnegie Mellon Egyetemen alakult Allen Newell és Herbert Simon közreműködésével. Az első számítógépekről szóló konferenciát 1956-ban Dartmouth-ban tartották, melyen azt kívánták modellálni, hogy hogyan gondolkodik az ember. Így tehát a játékok, mint például a sakkozás, vagy fordítások készítésére próbálták a gépeket programozni. A technika fejletlensége miatt nem sikerült megfelelő eredményeket felmutatni.

Jóllehet nem igen vagyunk tudatában, de szinte minden nap használjuk a mesterséges intelligencia kutatásainak eredményeit, amint bekapcsoljuk a számítógépünket. Lawrence Tesler úgy határozta meg a mesterséges intelligenciát a számítástechnikához viszonyítva, hogy *„minden, amit még nem csináltunk meg”* (McEnery, 1992: 10). E szerint a mesterséges intelligencia mai problémái holnapra már közhelyekké válnak. Ha valaki ki szeretne próbálni egy korai, de mai napig jól ismert programot, annak Weitzenbaum (1976) Eliza nevű programját javasoljuk. Ez a program a Turing tesztet is kiállja, ami abból áll, hogy miközben egy személy a számítógéppel és egy másik személlyel kommunikál a számítógépen, meg kell mondania, hogy mikor „beszél” a gép és mikor a másik személy. Az interneten a következő két címen lehet elérni a program egy változatát, és angol nyelven „társalogni” a számítógéppel: http://www.manifestation.com/neurotoys/eliza.php3 és http://www-ai.ijs.si/eliza/eliza.html.

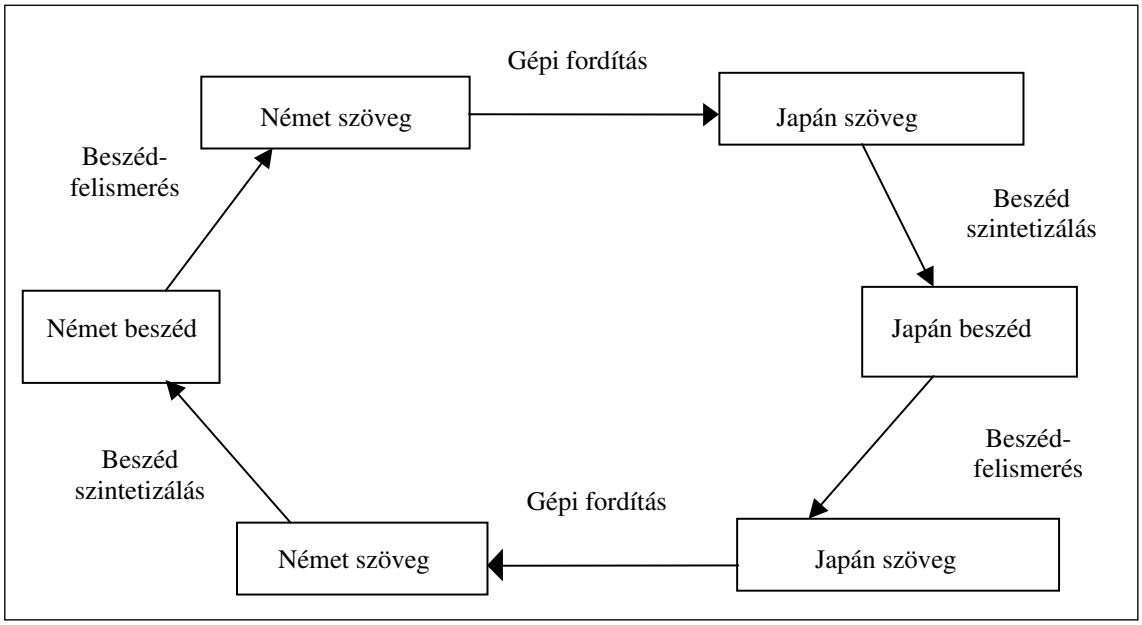
2.3.2.3. A számítógépes nyelvészet kutatási területe

Mint említettük, a számítógépes nyelvészet a nyelvészet és a mesterséges intelligencia interdiszciplináris kutatási területe, melynek műveléséhez mindkettő tudományág alapos ismerete szükséges. Ennek ellenére az 1950-es években a korai kutatásokat ezen a területen elsősorban mérnökök és számítógépes szakemberek végezték, akik nem sokat tudtak a nyelvről vagy nyelvészetről. Figyelmüket elsősorban a számítógépes fordítás kötötte le, melyről azt hitték, hogy viszonylag egyszerűen megoldható probléma. Azt képzelték, ami a mai nyelvtanulók körében sem ritka, hogy elegendő pusztán összepárosítani az egyik nyelv szavát a másik nyelv szavával, és kész a jó fordítás. Így fordulhatott elő, hogy a következő angol modatot: *“The spirit is willing but the flesh is weak.”* (ford. A szellem hajlandó, de a test gyenge, Máté 26, 41) oroszra fordították, majd ismét angolra és a következő mondat lett belőle: *“The vodka is excellent but the meat is rotten”* (ford. A vodka kiváló, de a hús rohadt.) (Pfaffenberger, 1993: 41). Nyilvánvaló, hogy ha egy adott szó több fordítási megfelelővel is rendelkezik, akkor nem lehet véletlenszerűen kiválasztani, mely fordítását használjuk. Ezt a fenti példa is bizonyítja.

A számítógépes nyelvészek ma is foglalkoznak a gépi fordítással, de nem tekintik azt kizárólagos feladatuknak. Kutatásaik többek között a következőket is vizsgálják: 1) természetes nyelvi interfész, 2) helyesírást ellenőrző programok és szókincstárak, 3) dokumentum-feldolgozás, 4) stílusellenőrzők és 5) lexikográfia.

A tágabban értelmezett számítógépes nyelvészet a beszédfelismeréssel is foglalkozik. A beszédfelismerő programok használatakor gondot okozhat a különböző hangok felis-

merése, például férfi-női, vagy felnőtt-gyerek hang, ezért e programok sokszor kiválóan működnek egy bizonyos beszélő vagy beszélő típus esetében, míg mások számára teljesen használhatatlanok. Így tehát azok a programok működnek megfelelően, amelyek vagy „taníthatók" a beszélőhöz való alkalmazkodásra, vagy eleve nem beszélőfüggők. 2000 augusztusában Hermann Maurer (személyes közlés, Tokyo) megemlítette, hogy a német Siemens cég szerint a közeljövőben egy német személy telefonbeszélgetést folytathat egy japánnal anélkül, hogy bármelyikük is beszélne a másik nyelvén. A következő ábra szemlélteti, hogy ez hogyan lehetséges.



18. ábra: A technológiával támogatott valós időben történő kommunikáció

Természetesen bármilyen nyelvvel helyettesíthető a japán és a német, csak megfelelő minőségű beszédfelismerő és szintetizáló programra van szükség az adott nyelven. A gépi fordítórendszernek gyorsnak és pontosnak kell lennie. Először is képesnek kell lennie arra, hogy az azonos hangzású szavakat helyesen értelmezze és írja le, vagy az azonos alakú, de több jelentésű szavak esetében a fordításhoz a megfelelő jelentést válassza ki. Ez a rendszer teljesen automatikusan, emberi közreműködés nélkül működne, és semmilyen közbeeső nyelv, mint például az angol nem szerepel. Nyilvánvaló, hogy egy közbülső természetes nyelv csak fokozná a hibalehetőségek számát. Hogy miért éppen a német–japán rendszer van már jelenleg kísérleti stádiumban? Nyilván gazdasági és technikai okok játszanak ebben döntő szerepet. Így hát ne is nagyon várjuk, hogy mi, magyarok is cseveghetünk majd másokkal nyelvtanulás nélkül.

2.3.3. A magyarországi számítógépes nyelvészetről

Mindenkiben felmerülhet a kérdés, hogy vajon Magyarországon mikor kezdtek el a számítógépes nyelvészettel foglalkozni? A válasz talán meglepő lesz. Már 1961-ben

megkezdődtek ilyen jellegű kutatások, és 1963-tól kezdve A számítógépes nyelvészet című szakkiadványt is kiadták évente egyszer vagy kétszer. A kiadvány 500 példányban jelent meg, a szerkesztőbizottság tagjai Papp Ferenc, Petőfi S. János, Szépe György és Varga Dénes voltak. 1975-re nem csak a szocialista blokk országaiban, hanem a John Benjamins B. V. kiadó révén az egész világon elérhetővé vált a *Computational Linguistics and Computer Languages* ('Számítógépes nyelvészet és számítógépes nyelvek') címmel. Tudomásom szerint az utolsó szám 1982-ben jelent meg.

Papp Ferenc (1930–2001) a Debreceni Egyetemen (amelyet akkor még Kossuth Lajos Tudományegyetemként ismertünk) állította be az első számítógépet a humán tudományok szolgálatára. Hunyadi László a magyar számítógépes nyelvészetről írt cikkében (1999) Papp Ferenc úttörő munkásságának egyik nagy eredményeként megemlíti *A magyar nyelv szóvégmutató szótárát* (1969). Papp Ferenc magyar, angol és orosz nyelven írt és szerkeszett könyvei, mint például a *Matematikai nyelvészet és gépi fordítás a Szovjetunióban* (Papp *et al.*, 1964), vagy a hasonló címmel angolul megjelent könyve (1966) mutatja, hogy milyen korán megindult Magyarországon az érdeklődés a számítógépes nyelvészet iránt.

A hazai számítógépes nyelvészet ismertetése során ki kell térnünk az MTA Nyelvtudományi Intézete Fonetikai Osztályán az 1980-as években nemzetközi szinten is úttörő beszédszintetizálási kutatásokra. E munkálatok keretében magyar és orosz nyelvű szöveg–beszéd (text to speech) rendszereket fejlesztett ki Bolla Kálmán, Kiss Gábor, Nikléczy Péter és Olaszy Gábor.

Kiefer Ferenc akadémikus a másik prominens személyisége a hazai számítógépes nyelvészetnek. 1992-től 2002-ig igazgatta a Magyar Tudományos Akadémia Nyelvtudományi Intézetét, amely mindig is kulcsszerepet játszott a hazai nyelvészeti kutatások fejlődésében. Kiefert már képzettsége is predesztinálta a számítógépes nyelvészettel való foglalkozásra, hiszen nyelvi diplomái mellett matematikusi diplomát is szerzett. Már 1964-es első publikációi is (Kiefer, 1964a, 1964b) azt jelezték, hogy elsősorban a nyelv és a nyelvészet érdekli, a matematikát pedig eszköznek tekinti, amely a nyelv megismerésében segíti. Számos írása jelent meg nemcsak itthon, hanem külföldön is, többek között a *Mathematical Linguistics in Eastern Europe* (1968) című könyve. Kiefer vezetése alatt az intézetben számos nemzetközi tudományos kutatás indult meg, melyek egy részét az Európai Unió támogatja. Kiefer Ferenc tiszteletére barátai és tanítványai könyvet adtak ki, melynek előszava (Szépe, 2001) bővebben ismerteti életútját és munkásságát.

Jóllehet a mai napig nincs önálló számítógépes nyelvészeti osztálya a Magyar Tudományos Akadémia Nyelvtudományi Intézetének, nemzetközileg is elismert szinten folyik az ilyen jellegű kutatómunka Váradi Tamás vezetésével a Korpusznyelvészeti Osztályon. Annak ellenére, hogy hivatalosan csak 1997 óta működik ez az osztály, a nyelvtechnológiai kutatások és fejlesztések már évekkel korábban megkezdődtek. A Nyelvtudományi Intézetben 1985-ben kezdődött el egy projekt, melynek az volt a célja, hogy a *Magyar nagyszótár* szerkesztését egy magyar nyelvű korpusz létrehozásával segítse. Pajzs Júlia a projekt előrehaladásáról és különböző aspektusairól számol be cikkeiben (pl., 1990, 1991, 1994; Pajzs *et al.*, 1992).

Az eddigiekből is kitűnik, hogy az e témával foglalkozó írások nagy része külföldön, vagy ha itthon is, de idegen nyelven jelent meg. Magyar nyelven három könyv jelent

meg (Bach, 1995; Prószéky, 1989; Prószéky & Kis, 1999). Bach írása az elektromérnöki kar hallgatói számára készült tankönyv, így nem széles körben terjesztik, és elsősorban a matematikai nyelvészet gyakorlati oldalával foglalkozik.

Prószéky Gábor könyve, *Számítógépes nyelvészet – Természetes nyelvek használata számítógépes rendszerekben* (1989), elsősorban a természetes nyelv feldolgozásával (NLP – natural language processing) és a természetes nyelv megértésével (NLU – natural language understanding) foglalkozik. A közel 600 oldalas könyvben négyoldalas (489–492) összefoglalást is találunk a számítógépes nyelvészet történetéről, mely a magyar nyelv feldolgozásáról szóló V. részt (489–550) vezeti be.

Prószéky és Kis *Számítógéppel emberi nyelven – Intelligens szövegkezelés számítógéppel* (1999) című könyve a szövegfeldolgozással kapcsolatos tudnivalókat írja le hétköznapi nyelven. E mellett azonban háttér információval is szolgál a gépi fordítás, lexikográfia és általában a számítógépek nyelvészetben való felhasználásának jobb megértéséhez. E könyvben számos olyan program működését is ismertetik, melyet a Prószéky által vezetett cég, a MorphoLogic készített, és a magyar nyelvű szövegfeldolgozó programok részeként működnek. A HuMor (Prószéky & Tihanyi, 1993) nevű program különösen figyelemre méltó, hiszen morfológiai elemző, amely elengedhetetlen része a magyar helyesírást ellenőrző programnak. Érdemes itt megemlíteni, hogy a cég honlapjáról elindulva http://www.morphologic.hu/h_pub.htm számos tudományos cikket találhatunk a linkeket követve. A címek már önmagunkban hasznos kis irodalomjegyzékként használhatók, hiszen Magyarországon nincs olyan jellegű szakfolyóirat, amely kizárólag a számítógépes nyelvészettel foglalkozna. Sajnálatos volt azonban, hogy a 2002-ben megjelent cikkek után semmilyen publikációt nem említettek a honlapon egy-két éven keresztül, még csak listaszerűen sem, és ez jelentősen megnehezítette írásaik figyelemmel kísérését. Örvendetes azonban, hogy Prószéky Gábor a közelmúltban felfrissítette publikációinak listáját, így már néhány 2005-ben megjelent írásának adatait is megtalálhatjuk itt.

A Nyelvtechnológiai Csoport 1998-ban a József Attila Tudományegyetem (2000 januárjától Szegedi Tudományegyetem) Informatikai Tanszékén alakult meg. A magyar nyelvre vonatkozó nyelvtechnológiai kutatások nagyon intenzíven folynak, és már szép eredmények is születtek. A különböző projektek sikere érdekében konzorciumot alkottak a fent említett MorphoLogic Kft. Budapest és az MTA Nyelvtudományi Intézet Korpusznyelvészeti Osztályával. A projektek ismertetését a csoport honlapján olvashatjuk http://www.inf.u-szeged.hu/projectdirs/hlt/. A honlap dicséretes részletességgel szól a kutatásokról és azok előzményiről is, így ezen a téren a legtöbb információt innen tudhatja meg az érdeklődő.

A Szegedi Tudományegyetem Informatikai Tanszékcsoportjának nevéhez fűződik a Magyar Számítógépes Nyelvészeti Konferencia megrendezése is. Az első konferencia 2003-ban, a második pedig 2004-ben tette lehetővé a hazai szakembereknek, hogy legfrissebb kutatásaik eredményeit bemutathassák és megvitathassák. Remélhetjük, hogy egy minden évben megrendezésre kerülő konferencia fogja a jövőben segíteni a hazai szakemberek munkáját és eszmecseréjét.

2.4. Folyóiratok

Mint említettük, a magyar nyelvű szakirodalom e téren könyvekben is igen szegényes, és a diszciplínát képviselő rendszeresen megjelenő folyóirat sem létezik, így néhány külföldi folyóiratot szeretnénk az érdeklődő olvasók figyelmébe ajánlani. A folyóiratok angol nyelvűek, de ha az egyes szerzők nevét egy internetes vagy könyvtári katalógusban megkeressük, más nyelven írott cikkekhez is eljuthatunk. Az első kettő után felsorolt folyóiratoknak nem minden száma tartalmaz feltétlen a számítógépes nyelvészethez kapcsolódó cikkeket. Esetleg több évfolyamot is át kell böngészni, mire az érdeklődésünknek megfelelőt megtaláljuk. Az itt felsorolt folyóiratok legtöbbjét a http://www.ingentaconnect.com/ honlapról elindulva elérhetjük, de a kiadóik honlapján is megtalálhatók. A folyóiratok tartalomjegyzéke és az összefoglalások ingyenesen megtekinthetők, csak a teljes cikkek letöltéséért kell fizetni, ha a könyvtárnak nincs előfizetése az adott folyóiratra.

- A **Computational Linguistics** című folyóirat (ISSN 0891-2017) az egyetlen, amelyet kizárólag a természetes nyelvfeldolgozásnak szenteltek. 1974-ben jelent meg az első szám, évente négyszer jelenik meg. Az 1994-től megjelent számok elektromos formában is megtekinthetők az előfizetők számára. A tartalomjegyzéket és a cikkek absztraktjait mindenki szabadon megtekintheti. A folyóirat honlapjáról http://mitpress.mit.edu/journals minden szám egy-egy cikkét ingyenesen is hozzáférhetővé tették. Egyes számokat kimondottan egy témának szenteltek, pl. a 29. évfolyam 3. száma (2003. szeptember) a „Web, mint korpusz” témával foglalkozik.
- **Journal of Machine Learning Research** (JMLR) on-line változata teljes egészében ingyenesen hozzáférhető a http://jmlr.org címről. Az első szám 2000 októberében jelent meg, egyes számok tematikusok, egy bizonyos témával foglalkoznak.
- **International Journal of Corpus Linguistics** (IJCL) a közelmúltig az egyetlen olyan folyóirat volt, amelyet teljes egészében a korpusznyelvészetnek szenteltek.
- **Corpus Linguistics and Linguistic Theory** teljesen új folyóirat, mely a http://www.degruyter.com/rs/384_7546_ENU_h.htm címen található meg. Első számának megjelenését 2005-re tervezik. A nyelvészet bármely területét érintő, korpusz alapú, de elméleti vonatkozású kutatások eredményeit kívánják itt publikálni.
- Applied Intelligence
- Artificial Intelligence
- Artificial Intelligence Review
- Computational Intelligence
- Computer Assisted Language Learning
- Computer Speech and Language
- Computers and Composition
- Computers and the Humanities
- Computers and Translation

- Hebrew Computational Linguistics
- International Journal of Computer Processing of Oriental Languages
- Language and Computation
- Literary & Linguistic Computing
- Machine Translation
- Research on Language and Computation
- Web Journal of Formal, Computational and Cognitive Linguistics http://fccl.ksu.ru

2.5. Összefoglalás

Ebben a fejezetben elsősorban a számítástechnika és maga a számítógép fejlődését tekintettük át, hiszen a korpusznyelvészet és a hozzá kapcsolódó társtudományok előrelépése szinte teljes egészében ezen alapult. A hardver fejlődésével (gyors memória és a merevlemez kapacitása, operációs rendszer stb.) e tudományterületek is „robosztusabbakká" váltak. A korpusznyelvészet és a számítógépes nyelvészet fejlődését a nyelvészeti elméleti háttér is igen befolyásolta. Chomsky és az őt követő, elsősorban amerikai (generatív) nyelvészet hatására az elméleti nyelvészet vált uralkodóvá. A nyelvhasználat vizsgálatát, és az empirikus kutatásokat sokan időpocsékolásnak tekintették.

A korpusznyelvészet tanulmányozása során elengedhetetlen, hogy a kapcsolódó tudományokról ne tájékozódjunk valamennyire. Sok cikkről igen nehéz megmondani, hogy vajon a korpusznyelvészet vagy a számítógépes nyelvészet címszó alá esik-e. A mesterséges intelligencia meghatározása után a számítógépes nyelvészet kutatási területét körvonalaztuk, amely az első korszakban szinte kizárólag a gépi fordítással foglalkozott.

A magyarországi számítógépes nyelvészet főbb képviselőinek – Papp Ferenc, Kiefer Ferenc, Prószéky Gábor, és az MTA Nyelvtudományi Intézete – munkásságát dióhéjban bemutattuk, majd a hazai szakirodalom hiányában néhány külföldi szakfolyóiratról és elérhetőségükről adtunk felvilágosítást a teljesség igénye nélkül.

3. A KORPUSZOKRÓL

3.1. Bevezetés

Megszokott dolog, hogy a korpuszok ismertetése a Brown Korpusszal kezdődik, hiszen ez volt az első elektronikus korpusz. E fejezet azonban két „előfutárt" mutat be elsőként, majd jelentőségüknél fogva az angol nyelvű korpuszokkal folytatjuk, hiszen történetileg is azokat hozták először létre. Ezek után megkísérelünk minél több más nyelvű korpuszt is bemutatni. Célunk a tájékoztatás, útbaigazítás, az olvasó első, önálló lépése megtételének elősegítése. A szakember számára bármely korpusz szakmai szempontból érdekes lehet, az érdeklődők számára azonban a számukra ismeretlen nyelvvel és kultúrával foglalkozók talán kicsit unalmasnak tűnhetnek. Ezért azt javasoljuk, hogy a korpuszok fejlődését bemutató rész után (3.9.-ig) az olvasó számára érdektelennek tűnő, az egyes nyelvek korpuszait ismertető részeket bátran ugorja át, és esetleg később térjen vissza a kihagyott részekhez.

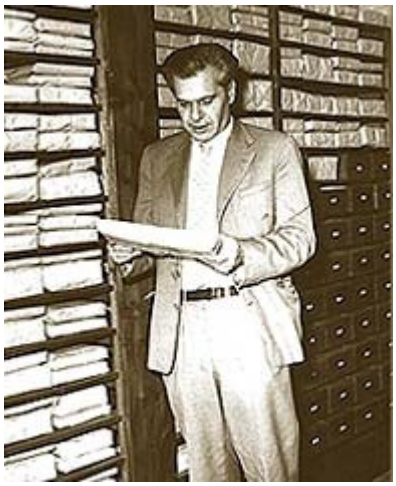
3.2. Az elektronikus korpuszok előfutárai

3.2.1. A Szerb Nyelv Korpusza

Annak ellenére, hogy a nemzetközi szakirodalomban és köztudatban csak a Survey of English Usage Corpust említik, amikor a modern, nem elektronikus korpuszok kerülnek szóba, vizsgálódásom eredménye alapján azt kell mondanom, hogy a Szerb Nyelv Korpusza megelőzi azt. Đorđe Kostić (1909–1995) az 1950-es évek elején foglalkozott a gépi fordítás, automatikus szöveg- és beszédfelismerés problémáival, és azt vallotta, hogy ezeket csak probabilisztikus (valószínűségen alapuló) módszerekkel lehet megoldani. Akkoriban ez nem volt divatos nézet, hiszen mindenkit inkább az algoritmusok, szabályrendszerek érdekeltek. A probabilisztikus módszerek alkalmazásához nagy mennyiségű szövegre van szükség, így az 1950-es évek közepén Đorđe Kostić megkezdte a korpusz létrehozását.

Az eredeti korpusz 11 millió (!) szóból állt, a 12. századtól Kostić koráig terjedő szövegeket tartalmazott. A korpuszban minden szót lemmatizáltak és a nyelvtani kódolás is befejeződött már 1962-re. A szavakat nem csak szófaj szerint, de morfológiai szempontból is elemezték, így a kódok száma eléri a kétezret. 1957 és 1962 között több mint 400 fő dolgozott a korpuszon, ebből 80 nyelvész vagy más szakember volt. Jóllehet nem elektronikus formában tárolták az adatokat, nagy előrelátásra vallott, hogy a nyelvtanra vonatkozó információkat egy hat számjegyből álló kóddal írták le. Az 1950-es évek végén a szintaktikai elemzést is megkezdték, de a 60-as évek elején a projekt abbama-

radt. A gépi fordítás tanulmányozása céljából nem csak szerb, hanem angol, német és francia szövegeket is annotáltak.



19. ábra: Đorđe Kostić a korpusz anyagával a háttérben (1959)

E korai korpuszelemzések eredményeinek egy részét Đorđe Kostić számos publikációban tette közzé az általa 1949-ben alapított intézet (Institute for Experimental Phonetics and Speech Pathology) kiadásában (Kostić, 1965a, 1965b, 1965c). E publikációk mellett, a korpuszra épülő különböző gyakorisági szótárak több mint 27 000 oldalt tesznek ki.

Mint említettük, a projektet 1962-ben megszakították, majd 1996-ban sikerült újraéleszteni. Első feladata az eredeti korpusz számítógépes feldolgozása volt, ami 1996 és 1999 között történt meg. A jelenlegi munkálatokról a későbbiekben, a fejezet más pontján fogunk szólni.

3.2.2. A SEU Korpusz (Survey of English Usage Corpus)

Randolph Quirk még a University of Durham angol nyelvészeti tanszékének professzora volt, amikor 1959-ben megalapította a Survey of English Usage[21]-ot (az angol nyelvhasználat felmérése), mely ma is a londoni University College-ban található. A korpusz készítését szociolingvisztikai céllal kezdték meg: a felnőtt, iskolázott brit lakosság nyelvtani és szóhasználati szokásait akarták vizsgálni. A terv egy egyenként 5000 szavas, 200 mintából álló korpusz létrehozása volt, mely el is készült. Így tehát a korpusz összesen egymillió szóból áll. A szövegek fele írott, a másik fele pedig beszélt nyelvi adatokat tartalmaz, melyek kissé formálisak és tudományosak. Az 8. táblázat mutatja a korpusz összetételét Kennedy alapján (1998: 17). A táblázatokban szereplő számok az adott csoportba tartozó szövegek számát jelentik.

[21] A Survey of English Usage a Londoni University College Angol Nyelv és Irodalom Tanszékén belül működő kutatási egység, mely az általuk készített korpusz neveként vált ismertté.

Nyomtatásban megjelent	Informatív (hírközlő)	Sajtó	8
		Tudományos	13
		Adminisztratív	4
		Jogi	3
	Oktató		6
	Meggyőző		5
	Kitalált		7
Nyomtatásban nem megjelent	Levelezés	Magán	13
		Nem magán	8
	Személyes naplók		4
	Folytatásos	Kitalált	5
		Informatív	6
Forgatókönyv alapján	Beszédek		6
	Drámai művek		4
	Hírek		3
	Előre megírt szónoklatok		3
	Történetek		2

8. táblázat: A SEU írott eredetű szövegei

Monológok			
Spontán			Előkészített, de nem megírt (6)
Szónoklatok (10)	Kommentárok		
	Sport (4)	Nem sport (4)	

Párbeszédek		
Személyes beszélgetések		Telefonbeszélgetések (16)
Titokban felvett (34)	Nem titokban felvett (26)	

9. táblázat: A SEU beszélt nyelvi szövegei

A SEU korpusz elemzése eredetileg kézzel és papíron történt, részletes nyelvtani annotációkat tartalmazott, amelyeket később számítógéppel feldolgoztak. Az anyaggyűjtés, átírás, feldolgozás több mint 25 évet vett igénybe (a szövegek 1953–1987 közöttiek), 1989-ben fejezték be. A korpuszt több mint 200 publikációhoz használták eredeti vagy számítógépes formájában, melyek listája a következő honlapon található: http://www.ucl.ac.uk/english-usage/archives/seu-biblio.htm.

Az 1970-es években a beszélt nyelvi szövegekből álló 500 000 szavas alkorpuszt számítógépes szalagra vették, ez a London–Lund Corpus (LLC). Ezt a korpuszt prozódiai annotációval is ellátták. A korpusz CD-ROM-on az International Computer Archive of Modern English-től (ICAME) szerezhető be.

3.3. A Brown Korpusz (1964)

A Brown Korpusz, teljes nevén Brown University Standard Corpus of Present-Day American English, a világ első elektronikus korpusza, mely 500 darab 2000 szóból álló szövegből áll, azaz 1 000 000 szövegszó a teljes korpusz (Francis & Kučera, 1964). Minden benne szereplő szöveg 1961-ben került kiadásra az Amerikai Egyesült Államokban. Az első fejezet 2–4. ábrája mutatja a korpusz összetételét és a benne szereplő különböző kategóriák korpuszon belüli arányát. Mivel ez volt az első, számos nyelvész követte a Brown Korpusz példáját, amikor saját korpuszukat megalkották. Ez részben idő- és energiatakarékosság miatt történt, hiszen egy korpusz létrehozása is rengeteg időt igényel, nemhogy kettőé. A nyelvészeti összehasonlító vizsgálatok elvégzéséhez pedig legalább két hasonló összeállítású korpuszra van szükség, így kézenfekvő volt egy már létező korpuszt felhasználni ehhez. A következő lista jól szemlélteti, hogy milyen nem amerikai angol nyelvhasználatot tükröző korpuszok készültek a Brown Korpusz mintájára:

- Lancaster–Oslo/Bergen Corpus (LOB) – brit angol 1961-ből
- Kolhapur Corpus of Indian English (KOL) (lásd Shastri, 1988) – indiai angol
- Freiburg–LOB Corpus (FLOB), brit angol az 1990-es évek elejéről
- Freiburg–Brown Corpus (FROWN), amerikai angol az 1990-es évek elejéről
- Australian Corpus of English (ACE), amelyet Macquarie Corpus of Written Australian English néven is emlegetnek
- Wellington Corpus of Written New Zealand English (lásd Bauer, 1993b) – újzélandi angol
- International Corpus of English (ICE) (lásd Greenbaum, 1992; Leitner, 1992b) – nemzetközi angol
- the Corpus of English-Canadian Writing – kanadai angol

A Kolhapur Corpus of Indian English és a Corpus of English-Canadian Writing nem követik teljesen a Brown Korpusz struktúráját. Számos más korpusz esetében is apróbb módosításokat kellett tenni. Például a Wellington Corpus of Written New Zealand English esetében a szokásos egyéves periódus helyett a szövegek egy 4 éves időszakból származtak.

A Brown Korpusz az egyik leggyakrabban elemzett korpusz, amelyet szintén az ICAME-től lehet beszerezni nem csak Key Word in Context (KWIC), azaz a keresett szóval a kontextusban, hanem szófaji és szintaktikai elemzési kódokat tartalmazó annotált változatban is.

3.4. A LOB Korpusz

Ez a korpusz egy nyolcéves együttműködés (1970–1978) eredményeképpen jött létre a Lancasteri és az Oslói Egyetem, valamint a Bergenben működő Norvég Társadalomtudományi Számítástechnikai Központ (Norwegian Computing Centre for the Humani-

ties) munkájának eredményeként. A korpusz létrehozásakor azt tartották szem előtt, hogy a Brown Korpusszal összehasonlítható, brit angol nyelvű korpuszt alkossanak, ezért a szövegeket ennek megfelelően a Brown Korpusz szövegeivel azonos évből, 1961-ből válogatták. A Johannson et al. (1978) által leírt különbségektől eltekintve a korpuszok gyakorlatilag azonos jellegűek és így összehasonlíthatók. A számítógép fejlődésének köszönhetően (a 4. és 5. generációs időszak) a kódolás gyorsabb és hatékonyabb volt.

A LOB Korpusz elemzésének eredményeit is számos cikk taglalja: Atwell, Leech és Garside (1984) a korpusz elemzését adja, Meijs (1984) a Brown és a LOB korpuszban szereplő elliptikus szerkezeteket hasonlítja össze, Tottie, Eeg-Olofsson és Thavenius (1984) a LOB és az LLC-n végzett negatív mondatok címkézéséről szól, Wikberg (1992) pedig diskurzus szempontjából vizsgálja meg a LOB és a Brown Korpuszt. Lásd még Atwell (1993), Biber (1990), és Valera & Rizo-Rodriguez (1998).

3.5. A COBUILD projekt

A COBUILD projekt a Birminghami Egyetem és a Collins Publishers nevű kiadó (később HarperCollins) együttműködésének eredményeképpen 1980-ban kezdte meg működését. Két fő célja volt: 1) nagy terjedelmű, számítógéppel feldolgozott modern angol nyelvű korpusz gyűjtése és elemzése; 2) az eredmények publikálása az angolt idegen nyelvként tanuló diákok és oktató tanárok számára készült referencia és oktató könyvek széles skáláját létrehozva (Krishnamurthy, 1997b). Jóllehet az „első jelentős számítógépes korpusz-alapú lexikográfiai projekt" az American Heritage Project volt (Kennedy, 1998: 61) az 1970-es években, amely így megelőzte a COBUILD projektet, de a korpusznyelvészet elterjedésében nem játszott olyan nagy szerepet, mint a COBUILD projekt. Ennek oka talán az, hogy a COBUILD projekt első eredményeként kiadott korpusz-alapú szótár, a *Collins COBUILD English language dictionary* (J. Sinclair, 1987a), az EFL (angol mint idegen nyelv) piacon szinte robbanásszerű változást hozott. Minden magára adó szótárkiadó követte példáját. Manapság még eladhatóak a nem korpusz-alapú szótárak, de a vásárlók egyre növekvő hányada számára a szótár anyagát adó korpusz a megbízhatóság záloga.

A projekt keretében számos korpuszt és alkorpuszt hoztak létre, és használtak a lexikográfiai vizsgálatokhoz. A korpusz tervezése és az engedélyek beszerzése 1980-ban kezdődött. Az adatbevitel 1981-ben indult meg. Sinclair a *Looking up: An account of the COBUILD project in lexical computing* (1987b) című könyvében részletesen beszámol a projektről.

Az első korpusz a Main Corpus (Fő Korpusz) volt, 7,3 millió szót tartalmazott. Ezt a Reserve Corpus (Tartalék Korpusz) követte 11 millió szóval 1985-ben (Renouf, 1987). A korpuszépítés azonban egy pillanatra sem állt meg, így az egyes alkorpuszok is folyamatosan bővültek, új alkorpuszokat hoztak létre. Ezért ha Birmingham nevét hallja az ember korpusznyelvészeti körökben, akkor manapság nem ezekre a korpuszokra gondol, hanem a Bank of English (Az angol nyelv tárháza) jut eszébe, amelyet 1991-ben indítottak útjára. A folyamatos hozzáadások révén 1993-ra már 120 millióra, 1994-

re 167 millióra, 1995-re pedig több mint 320 millió szóra növekedett ez a korpusz (Krishnamurthy, 1997a). A következő táblázat a Bank of English 1995-ös összetételét mutatja. A korpusz mindössze 27,47%-a származik nem brit eredetű forrásból: 22,62% amerikai és 4,85% ausztrál eredetű. A Bank of English jelenleg 524 millió szóból áll és állandóan növekszik. Ebből, az interneten keresztül, egy 56 millió szóból álló részt, a COBUILD Direct Corpus-t bárki számára, díj megfizetése mellett hozzáférhetővé tettek. Mind intézmények, mind magánszemélyek használhatják. Horváth József (1999) a Modern Nyelvoktatás című lapban ismertette magyar nyelven e szolgáltatást.

A Bank of English összetétele 1995 áprilisában

Forrás	Méret (millió szó)	Szövegek száma	A korpusz %-os aránya
BBC World Service	18,7	(500)	8,88%
Independent (napilap)	5,0	49	2,38%
Times (napilap)	10,3	79	4,89%
National Public Radio* (Washington)	22,0	729	10,45%
Economist (folyóirat)	8,7	28	4,16%
Szórólapok	1,8	837	0,86%
New Scientist (folyóirat)	4,1	92	1,95%
Ausztrál hírek	10,2	141	4,85%
Beszélt nyelv (általános)	15,5	1571	7,36%
Today	18,1	540	8,6%
Amerikai könyvek és újságok*	19,4	273	9,22%
Brit könyvek	27,9	406	13,25%
Guardian (napilap)	12,6	137	5,99%
Magazinok (általános, divatos)	30,0	760	14,25%
Wall Street Journal*	6,2	15	2,95
TOTAL	210,5	6156	100
További korpuszok:			
Üzleti angol nyelv (újságok és tankönyvek)	3,0		
Egyetemi/tudományos írások	1,5		
GCSE (középiskolai záróvizsga) tankönyvek	1,0		

A * amerikai angol nyelvet jelöl.

10. táblázat: A Bank of English szerkezete Krishnamurthy alapján (1997a: 79)

A COBUILD projekt célja nem csak referenciakönyvek kiadása volt, hanem a korpuszra épülő pedagógiai jellegű segédkönyvek és tankönyvek megjelentetése is. A Willis házaspár *Collins COBUILD Course of English* (1988) című tankönyvsorozata valóban eredetien közelítette meg az angolt idegen nyelvként tanuló diákoknak szóló tankönyv írását. Az egyes tankönyvekben szereplő tanításra szánt szavak kiválogatásakor a korpuszelemzések eredményeit, elsősorban az előfordulási mutatót vették figyelembe (lásd még D. Willis, 1990).

3.6. A Brit Nemzeti Korpusz – British National Corpus (BNC)

A Brit Nemzeti Korpusz számos intézmény és kiadó együttműködésének eredményeképpen jöhetett létre, melyek a következők: Oxfordi Egyetemi Kiadó (Oxford University Press), a Longman Csoport (Longman Group, UK), Chambers kiadó, a British Library (könyvtár), valamint az Oxfordi és Lancasteri Egyetem (lásd http://www.natcorp.ox.ac.uk/). Az együttműködés azért is nagyon szerencsés volt, mert minden intézmény olyan feladatokat végzett el, amelyekhez a legjobban értett. A korpusz készítésének megkezdése előtt igen komoly tervező munkát végeztek, hogy a korpuszba kerülő anyagok a lehető legjobban tükrözzék, azaz reprezentálják az angol nyelvet. Ez természetesen a helyes arányok megtartására való törekvést is jelentette. Ne feledjük azonban, hogy a nyelv teljes mennyiségére vonatkozóan nem is lehetnek pontos adataink, így a már említett kiegyensúlyozott jelző is csak becsült értékeket takar, és éppen ezért nem teljesen objektív.

A BNC 4124 szöveget tartalmaz, melynek 90%-a írott eredetű, és mindössze 10%-a származik a beszélt nyelvből. A reprezentativitás érdekében nagyon sok tényezőt kellett figyelembe venni a mintavételnél. A beszélt nyelvi korpusz esetében a beszélők lehető legszélesebb skáláját választották ki a következők alapján:

- kor szerint (6 csoportra osztott sávos megoszlás)
- nemek szerint (férfi és nő)
- társadalmi osztály/helyzet
- az ország területi megoszlása szerint
- monológ és párbeszéd (25%–75%)

Nyilvánvaló, hogy a mai magyar nyelvről sem kapnánk hű képet, ha például csak a zalaegerszegi 30–35 év közötti férfi orvosok párbeszédét gyűjtenénk össze és elemeznénk.

Az írott szövegek kiválasztása az időpont, a médium és a tartalom alapján történt. A szövegek 1960–1974 és 1975–1993 között születtek. A médium szerint a következő csoportosítást használták: könyv, folyóirat, egyéb nyomtatásban megjelent írás, egyéb nyomtatásban meg nem jelent írás, és beszédre szánt írás. Ha ezen csoportok egyikébe sem illett bele egy írás, akkor az „osztályozatlan” címszó alá került. A tartalom szerinti csoportosításkor a következő csoportokat hozták létre: szépprózа és informatív (hírközlő) próza, ez utóbbin belül pedig: természettudomány, alkalmazott tudomány, társadalomtudomány, nemzetközi ügyek, gazdaság, művészet, hit és gondolkodás, valamint szabadidő szerepelnek. A szerzőkre vonatkozó információk: nem, kor, lakcím, valamint hogy egyedüli szerzők-e vagy vannak társszerzőik. A megcélzott olvasóközönség nemét és korát is jelölik, valamint szocio-kulturális helyzetét magas, közepes vagy alacsony kategóriába sorolták.

A teljes korpuszt nyelvtani címkékkel látták el. Ezt a feladatot a Lancasteri Egyetem Angol Nyelv Számítógépes Kutatásának Csorportja (Unit for Computer Research on the English Language – UCREL) végezte az egyetemen kidolgozott CLAWS4 nevű automatikus címkéző program[22] (angolul: tagger) segítségével. Egy kétmillió szóból álló

[22] Az automata címkéző olyan számítógépes program, amely a korpusz minden szavához egy szófajt jelölő címkét illeszt.

rész címkézését egy ennél is pontosabban dolgozó, C7-ként ismert programmal is elvégezték, majd kézileg ellenőrizték és javították. Erre a kézi elemzésre azért volt szükség, mert az elemző program időnként nem tudja eldönteni, hogy egy adott esetben a több szófajként is előforduló szót hogyan jelölje, így előfordulhatnak hibák is, vagy egyszerűen hiányzik a jelölés. A hivatalos álláspont szerint a kétmillió szóból álló, a korpusz magjának (Core Corpus) nevezett rész elemzése esetében a hibaszázalék mindössze 0,3% volt. Ez a kézzel is ellenőrzött változat, valamint a teljes, 100 millió szóból álló BNC CD-ROM-on megvásárolható. A BNC-t díj ellenében az interneten keresztül is lehet használni. Bővebb információt a következő címről lehet megtudni: http://sara.natcorp.ox.ac.uk/

Mielőtt mindenki fejvesztve rohanna, hogy megvegye ezt az alacsony hibaszázalékkal elkészített CD-ROM-ot, azért gondoljunk a következőkre is. A nyelvészek egyáltalán nem értenek mindenben egyet az annotációt illetően. Magával a kódkészlettel sem, és esetleg az egyes szavak hovatartozását illetően sem. Így tehát a hibaszázalékot csak az adott címkerendszer keretében értelmezhetjük. Nos, ha valaki magával a kódolás rendszerével sem ért egyet, a hibaszázalék az ő szempontjából jelentősen megnőhet.

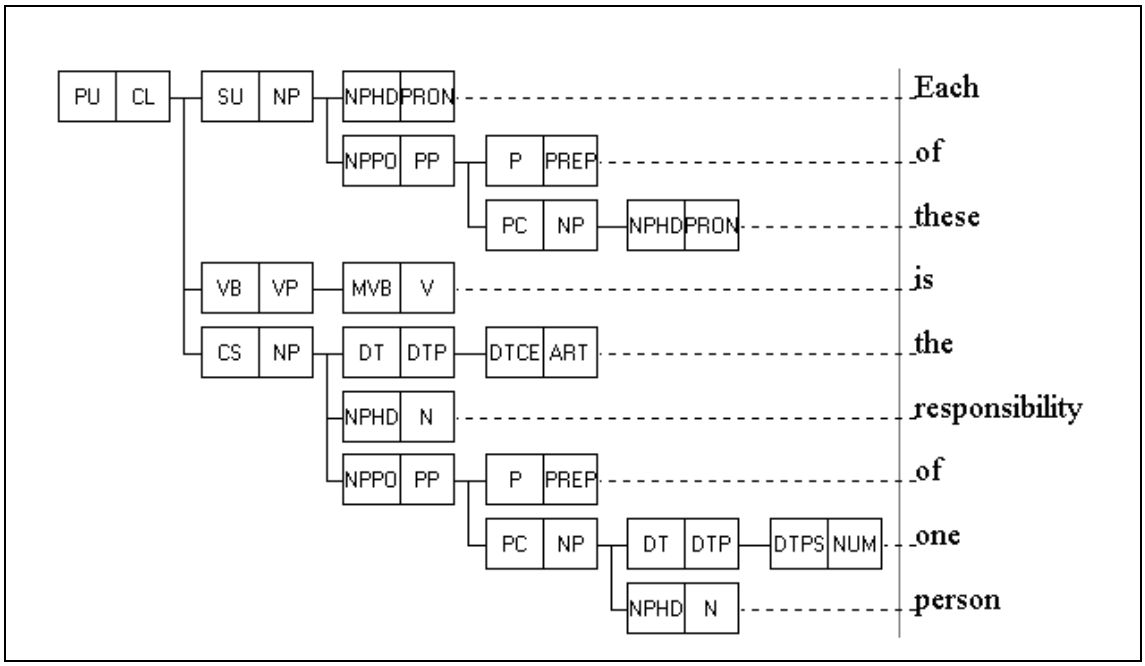
A további kritikák között említsük meg Lou Burnard előadását a 4. Nemzetközi Oktatás és Nyelvi Korpuszok Konferenciáról (Fourth International Conference on Teaching and Language Corpora – TALC), amelyben a BNC tervezéséről és bizonyos döntésekről szólt (L. Burnard, 2000). Már a cím is sokat sejtető: „Hol tévedtünk?" Burnard néhány rosszul meghatározott kategóriát kritizál, és azt is, hogy esetenként hiányzik a szövegre vonatkozó információ. David Lee (2000) egyenesen dzsungelnek nevezte a BNC-t, rámutatva arra, hogy milyen nehéz a BNC-ben eligazodni, és éppen ezért egy külön programot írt (BNC Indexer) arra a célra, hogy a kutatók gyorsan és könnyen megtalálhassák a kívánt szövegeket a zsáner és/vagy más szövegklasszifikációs kritérium alapján. A BNC-ben túlságosan átfogóak a kategóriák és sokszor félrevezetők a címek. Példaként azt említi, hogy az órai beszélgetések és a kiscsoportos szeminárium is az „előadás" címszó alatt szerepel, annak ellenére, hogy csak nagyon kevés résztvevője volt. Lee azzal is vádolja a BNC-t, hogy nem reprezentatív, hiszen néhány műfaj, mint például az orvosi vagy jogi tanácsadás, parlamenti viták, teljesen hiányoznak vagy csak igen kis arányban szerepelnek.

3.7. Az Angol Nyelv Nemzetközi Korpusza (International Corpus of English – ICE)

1988-ban Sydney Greenbaum azt javasolta, hogy hozzanak létre egy olyan nagyméretű korpuszt összehasonlító nyelvészeti célokkal, amely az angol nyelv összes változatát tartalmazza. Ez pedagógiai szempontból is igen fontos, hiszen még azok az anyanyelvi beszélők is, akik nem álltak kapcsolatban az angol más változataival a televízió, mozi vagy más révén, hamar rádöbbennek, hogy az angol anyanyelvi beszélők is sokszor „más nyelvet" beszélnek. Ez nem csak a kiejtésben vagy szókincsben, hanem a nyelvtani különbségekben is megnyilvánul. Ebben a korpuszban minden egyes alkorpuszt egymillió szóra terveztek, s mind beszélt, mind írott szövegeket tartalmaz. Azt is igen

pontosan meghatározták, hogy milyen nyelvi adatok kerülhetnek bele. Például a szerzőnek vagy beszélőnek 18 év feletti angol anyanyelvű beszélőnek kellett lennie, vagy olyan országban kellett felnőnie, ahol az angolt első idegen nyelvként vagy domináns nyelvként beszélték, és legalább a középiskola befejezéséig angol nyelvű oktatásban vett részt (Kennedy, 1998).

A szövegek az 1990–1996 közötti időszakból származnak. Az első elemzésre kész alkorpusz a brit angol volt. Az alkorpusz szerkezete, melyet a 2. táblázatban láthatunk, nagyon hasonlít a LOB és a Brown Korpuszéhoz, mivel itt is 500 darab körülbelül 2000 szóból álló szöveg alkotja a korpuszt. A legfőbb különbség az, hogy a LOB és a Brown Korpusz szövegei mind írott szövegek voltak, az ICE esetében viszont az írott és beszélt szövegek aránya 60% és 40%, mivel 300 szöveg beszélt nyelvi eredetű. (A SEU esetében 50–50% az arány.) Az ICE tanulmányozására külön számítógépes programot dolgoztak ki. Mint azt már a korpusz méretének tárgyalásakor (1.3.1.2. rész) említettük, jóllehet egy kisebb méretű korpusz nem alkalmas az alacsony előfordulású szavak vagy a lexikográfiai vizsgálatok elvégzéséhez szükséges információk szolgáltatására, hasznos lehet a nagy gyakorisággal előforduló szavak, így a prepozíciók vagy a nyelvtani tulajdonságok vizsgálatánál. A korpuszt nem csak a szövegre vonatkozó információkkal, hanem szófaji címkékkel és a mondattani elemzés címkéivel is ellátták. Az ICE honlapja szerint a szófaji elemző program pontossága 95% körül van. A nyelvtani elemzést végző programot a Nijmegeni Egyetem TOSCA nevű csoportja fejlesztette ki. Minden mondatot három szinten elemeznek: a szószerkezetek (phrase), a tagmondatok (clause) és a mondat (sentence) szintjén. Az eredményt egy ágrajz mutatja.



20. ábra: Mondattani elemzés ágrajza az ICE-ben (http://www.ucl.ac.uk/english-usage/ice/annotate.htm)

Az ICE brit angol korpuszából 10 szöveg, azaz 20 000 szó, valamint a teljesen működőképes elemző program ingyenesen letölthető a következő címről: http://www.ucl.ac.uk/english-usage/ice-gb/sampler/download.htm. Amennyiben a teljes korpuszra van szükség, az CD-ROM-on megrendelhető díjfizetés ellenében. Az alábbi táblázat a teljes brit korpusz összetételét mutatja. A jobb oldali oszlopban az ingyen letölthető szövegek mellett a fájl neve látható.

Beszélt nyelvi szövegek (300)	párbeszéd (180)	**magán (100)**	szemtől-szembeni beszélgetések (90) telefonhívások (10)	S1A-010 S1A-094
		nyilvános (80)	iskolai tanórák (20) közvetített viták (20) közvetített interjúk (10) parlamenti viták (10) jogi keresztkérdések (10) üzleti tárgyalások (10)	
	monológ (100)	**nem megírt (70)**	spontán kommentárok (20) megíratlan beszédek (30) demonstrációk (10) jogeseti előadások (10)	S2A-011
		előre megírt (30)	közvetített beszélgetések (20) nem közvetített beszédek (10)	S2B-026
	vegyes (20)		hírközvetítések (20)	S2B-002
Írott szövegek (200)	nyomtatásban nem megjelent (50)	**nem munkahelyi írások (20)**	házi diák esszék (10) vizsgai diák munkák (10)	W1A-001
		levelezés (30)	kapcsolattartó levelek (15) üzleti levelek (15)	W1B-001
	nyomtatásban megjelent (150)	**tudományos írások (40)**	humán (10) társadalomtudományos (10) természettudományos (10) technikai (10)	W2A-005
		nem tudományos írások (40)	humán (10) társadalomtudományos (10) természettudományos (10) technikai (10)	
		riport (20)	újsághírek (20)	W2C-009
		utasító jellegű írások (20)	adminisztratív/utasító (10) készség/hobbi (10)	W2D-018
		meggyőző írások (10)	vezércikk (10)	
		kreatív írások (20)	regény/novella (20)	

11. táblázat: Az ICE brit alkorpuszának összetétele (http://www.ucl.ac.uk/english-usage/ice-gb/sampler/download.htm alapján)

3.8. A nem anyanyelvi angol korpuszok

Az eddig említett korpuszokat elsősorban az angol nyelv pontosabb leírására, nyelvhasználatának elemzésére hozták létre. Nyelvészek és lexikográfusok munkájának elősegítése céljából készültek. Természetesen a pontosabb nyelvészeti elemzés, valamint

az, hogy a korpuszok példák százait vagy ezreit kínálják a nyelvtanároknak, sokat segített a nyelvoktatásban is. A COBUILD projekt keretében a nyelvoktatás elősegítése kimondott cél volt, és számos kiadvány jelent meg a diákok számára készült szótáron kívül is. Pedagógiai szempontból azonban nem csak a követendő példa megmutatása fontos, hanem azt is tudnia kell egy jó pedagógusnak, hogy a tanulás folyamatában milyen nehézségekkel küzdenek a diákok, milyen hibákat ejtenek. Az angolt idegen nyelvként tanuló diákokat kiszolgáló kiadóknak és egyéb oktatási intézményeknek így tehát információra van szüksége ahhoz, hogy még jobban igazodhassanak a diákok – országonként és szintenként – eltérő igényeihez. Elsősorban az egynyelvű tanulói szótárak kiadói rendelkeztek már eleve a korpuszkészítéshez szükséges számítógépes és szakemberi háttérrel, így nem véletlen, hogy a nem anyanyelvi korpuszok létrehozásában is élen jártak. A nagy lexikográfiai projektek, mint a COBUILD, nemcsak referenciakönyveket készítettek a nem anyanyelvi beszélők számára, hanem egyre több nyelvtanárral is megismertették az ezekben a projektekben használt módszereket. Ennek eredményeképpen mind több tanár úgy érzi, hogy a nem anyanyelvi beszélők produktumaiból készített korpusz is igen sokat segíthet a tanítás eredményességét illetően, hiszen fontos információkkal szolgálhat a helyesen és helytelenül használt nyelvtani vagy szókincsbeli, esetleg szövegszerkesztési hibákról. A nyelvtanárok segítsége nélkül nem lehetett volna ilyen korpuszokat létrehozni.

3.8.1. A Longman Angol Nyelvtanulói Korpusz – Longman Corpus of Learners' English (LCLE)

A Longman Angol Nyelvtanulói Korpusz, melyet mostanában csak Longman Learners' Corpus-ként emlegetnek, részét képezi a Longman Corpus Networknek, mely jelenleg öt részből áll, s melynek többi komponenséről később szólunk. Az LCLE kb. 10 millió szóból áll, és azzal a céllal készült, hogy segítse a tudományos kutatást, valamint a lexikográfiai és más oktatási jellegű művek kiadását (Warren, 1992). A korpusz 8 különböző tudásszintű, 160 különböző nyelvi háttérrel rendelkező diák szövegeit tartalmazza. A korpusz honlapját a következő címen találjuk: http://www.longman-elt.com/dictionaries/corpus/lclearn.html. A teljes korpusz lehetőséget nyújt arra, hogy fény derüljön olyan tipikus problémákra, melyek a nyelvtanulók anyanyelvétől teljesen függetlenek. Az egyes alkorpuszok vizsgálata pedig az anyanyelvfüggő tipikus hibákat mutatja meg. Az alkorpuszok között összehasonlító vizsgálatokat is lehet végezni. Természetesen ilyen jellegű kutatást bármilyen hasonló összeállítású korpuszon végezhetünk.

3.8.2. A Nemzetközi Angol Nyelvtanulói Korpusz (International Corpus of Learners' English (ICLE))

Ez a korpusz különböző nemzetiségű, haladó szinten álló nyelvtanulók írott szövegeinek gyűjteménye. Jelenleg 19 alkorpuszból áll, melyek egyenként 200 000 szót tartalmaznak. Minden diák maximum 1000 szóval szerepelhet a korpuszban, így az alkorpuszokban legkevesebb 200 diák írása található. Az esszékre, valamint az adatgyűjtésre

vonatkozó információkat a projekt honlapján találhatjuk: http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm.

Alkorpusz eredete	**Intézmény**	**ICLE honlap**	**A korpusz állapota**
Bulgária	Sofia University		teljes
Brazilia	Catholic University in Sao Paulo (PUC-SP) University of Sao Paulo (USP)	http://lael.pucsp.br/corpora/bricle/	
Kína	Lingnan University of Hong Kong Hong Kong Polytechnic University The Chinese University of Hong Kong		
Cseh Köztársaság	Jan Evangelista Purkyne University	http://kvt.ujep.cz/~flaskaj/icle/	teljes
Hollandia	Katholieke Universiteit Nijmegen		teljes
Finnország	Åbo Akademi University Växjö University	http://www.abo.fi/fak/hf/enge/iclefin.htm	teljes
Franciaország	Université catholique de Louvain	http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm	teljes
Németország	Universität Augsburg	http://wwwanglistik.phil.uni-augsburg.de/lorenz/	teljes
Olaszország	Università di Torino		teljes
Japán	Showa Women's University		teljes
Litvánia	Vilniaus Universitetas		
Norvégia	University of Oslo		teljes
Lengyelország	Adam Mickiewicza University (Poznan)	http://main.amu.edu.pl/~przemka/#PICLE	teljes
Portugália	Escola Superiro de Technologia de Viseu – Northumbria University University Nova de Lisboa University of Aveiro		
Oroszország	Lomonosov Moscow State University		teljes
Spanyolország	Universidad Complutense de Mardid		teljes
Dél-Afrika (Setswana)	Potchefstroom University		
Svédország	Lund University Göteborg University Växjö University Lund University	http://www.englund.lu.se/research/corpus/corpus/swicle.html	teljes
Törökország	Cukurova University		

12. táblázat: Az ICLE alkorpuszai (http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm alapján)

Sylvian Granger, a Louvaini Katolikus Egyetem tanára indította útjára a projektet, és számos publikációban tette közzé kutatásai eredményeit (1993, 1994, 1996). Mint lát-

hatjuk, magyar diákok írása nem szerepel a listán, így ha valaki kedvet érez az együttműködéshez, hozzájárulhat egy nemzetközi kutatás további sikeréhez.

3.8.3. A Hongkongi Műszaki és Természettudományi Egyetem Angol Tanulói Korpusza (Hong Kong University of Science and Technology [HKUST] Corpus of Learner English)

A HKUST nem csak egy korpusszal büszkélkedhet, ezért tartjuk fontosnak, hogy a félreértések elkerülése végett erre felhívjuk a figyelmet. A tanulói korpusz mellett öt témakörben angol nyelvű tankönyvek felhasználásával egyenként kb. 1 millió szavas korpuszokat hoztak létre (lásd Fang, 1992; Kam-mei *et al.*, 2003). A tanulói korpusz, melyet 5-6 millió szóra terveztek, a világon az egyik legnagyobb abban a tekintetben, hogy azonos anyanyelvű beszélők által írt nyelvtanulói szövegeket tartalmaz. Az elemzések eredményeit számos cikkben ismertették az egyetem oktatói.

3.8.4. Japán diákok angol nyelvű korpuszai

A japán oktatási minisztérium egyre nagyobb hangsúlyt fektet az eredményes iskolai nyelvtanításra, így nem véletlen, hogy Japánban is egyre többen kezdenek érdeklődni nem csak a „klasszikus” anyanyelvi korpuszok iránt, hanem a tanulók nyelvi produktumaiból létrehozott korpuszok iránt is. Tono Yukio munkássága a legismertebb ezen a területen, aki egyszerre több korpusz készítésével is foglalkozik. Az egyik korai korpusza, melyet doktori disszertációjához használt, 700 000 szóból állt. Ennek segítségével vizsgálta a diákok által a kollokációk terén elkövetett hibákat. Ha a hibákat feltárjuk, több figyelmet lehet fordítani a tanítás, a referencia- és tankönyvek készítése során ezek pontosabb bemutatására és használatára. Tono japán diákok írásait használta, és elemzései arra is rámutattak, hogy a hibák nagy része az anyanyelv (L1) hatásából eredt. Ez igen erős jelzés arra, hogy a kétnyelvű szótárak készítésekor a kiadóknak feltétlenül rendelkezniük kell saját tanulói korpusszal, hogy az elemezések eredményeit figyelembe vehessék a szótár vagy más referencia jellegű kiadvány készítésekor. Véthetnek ugyan hasonló hibákat különböző anyanyelvű diákok is – különösen, ha anyanyelvük azonos nyelvcsaládba tartozik –, de kevés a valószínűsége annak, hogy igen különböző nyelvi háttérrel rendelkező diákok, például japán és német diákok, nagyrészt hasonló hibákat kövessenek el. Tono honlapján http://leo.meikai.ac.jp/~tono/ számos tanulói korpuszról tesz említést, elsősorban japán diákok korpuszairól, amelyek közül egyesek letölthetők.

3.8.5. A Janus Pannonius Tudományegyetem Korpusza

Utolsónak említjük, de számunkra talán a legfontosabb, hogy magyar diákok írásaiból is készült tanulói korpusz. Bizonyára több korpusz is létezik, de kettőről van tudomásunk, melyek egyetemisták esszéit tartalmazzák. Horváth József, a Pécsi Jannus Pannonius Egyetem oktatója egy 412 280 szavas korpuszt hozott létre diákjai írásaiból (JPU Corpus), melyről magyarul is olvashatunk a Modern Nyelvoktatás című folyóiratban

(Horváth, 2000). Mindannyian tapasztalhattuk az évek során, hogy nem csak országok között vannak különbségek az oktatás területén, de még egy országon belül is jelentősek lehetnek az eltérések. Sőt, akár két iskola között is. Ezért tartotta fontosnak Horváth, hogy saját diákjainak írásaiból hozzon létre tanulói korpuszt, melynek elemzésével tisztábban láthatja diákjainak orvoslásra szoruló problémáit. A korpusz pedagógiai annotációjával külön cikkben foglalkozik (Horváth, 2002). Mivel Horváth ezt a korpuszt használta doktori dolgozata megírásához, az érdeklődők könyv formájában is olvashatnak vizsgálatainak eredményeiről (2001).

3.8.6. Az Eötvös Loránd Tudományegyetem Korpusza

Tankó Gyula, az Eötvös Loránd Tudományegyetem diákjainak vizsgafeladatként írt esszéit gyűjtötte össze és használta összehasonlító vizsgálatokra. A 93 darab, egyenként kb. 500 szavas esszét a fogalmazáson belüli kötőelemek vizsgálatának céljára használta fel. Kutatásának eredményeit angol nyelven a *How to Use Corpora in Language Teaching* (J. M. Sinclair, 2004b) című kötetben olvashatja a kutató tollából az érdeklődő.

3.9. A korpuszok nyelvenként

Az eddig említett korpuszok történeti jelentőségük vagy jellegük miatt kerültek a fejezet elejére, és ezek bemutatásával általános képet igyekeztünk nyújtani. A korpuszok száma azonban napról napra növekszik, méretük változik, így lehetetlen lenne teljes áttekintést adni az összes létező korpuszról. Soknak ugyanis még a létezéséről sem biztos, hogy tudunk. A fent említett, Horváth és Tankó által készített korpuszokon kívül is bizonyára készültek vagy készülőben vannak más korpuszok hazánkban is, amelyekről nincs tudomásunk. Így ne lepődjön meg az olvasó, ha a fejezet további részében leírt korpuszokon kívül másokat is felfedez kutatásai során, vagy az esetleg itt megadott adatoktól eltérőekkel találkozik. Azt is meg kell említenünk, hogy sok esetben átfedések vannak a különböző korpuszok között, így egy adott szövegcsoportot esetleg két különböző név alatt is megtalálhatunk. Például a következő szakaszban szereplő Longman Brit Beszélt Nyelvi Korpusz a Brit Nemzeti Korpusz beszélt nyelvi részeként is szerepel, a kettő teljesen azonos. Megpróbáljuk az ilyen esetekre a figyelmet felhívni.

Mivel az angol nyelvű korpuszok jelentős túlsúlyban vannak és a többi nyelv korpusza is az ezekre vonatkozó tapasztalatok felhasználásával készül, először az angol nyelvű korpuszokat ismertetjük röviden. Ezt követik a német és a francia nyelvű korpuszok. A környező országok nyelveit Magyarországon is sokan beszélik és tanulják, így a magyar nyelvű korpuszok ismertetése után ezek következnek. Az Európában beszélt nyelvek korpuszait néhány „egzotikus” nyelv követi, elsősorban a japán.

Az egyes korpuszokhoz kapcsolódó kereső programok bemutatására nem kerül sor, vagy csak ritkán. Az adott nyelvet beszélő érdeklődők számára az interneten a korpusz és a kereső használatára vonatkozó minden szükséges információ megtalálható az adott nyelven, így sok esetben csak a honlapok címeinek megadását tartottuk fontosnak.

3.9.1. További angol nyelvű korpuszok

3.9.1.1. A Brown Korpusz klónjai

- Lancaster–Oslo/Bergen Corpus (LOB)
- Kolhapur Corpus of Indian English (KOL) (Shastri, 1988)
- Freiburg–Brown Corpus (FROWN) és Freiburg–LOB Corpus (FLOB)
- Australian Corpus of English (ACE), amelyet Macquarie Corpus of Written Australian English néven is emlegetnek
- the Wellington Corpus of Written New Zealand English (Bauer, 1993a)
- the International Corpus of English (ICE) (Greenbaum, 1992; Leitner, 1992a)
- the Corpus of English-Canadian Writing

A Brown Korpuszról szóló részben (3.3.) már említést tettünk arról, hogy számos korpusz készült a Brown Korpusz mintájára, hiszen az azonos szerkezet lehetővé tette az összehasonlító vizsgálatok elvégzését. A LOB Korpuszról (3.4.) és az ICE-ről (International Corpus of English) (3.7.) külön is szóltunk, valamint az (ICE) szerkezetét össze is hasonlítottuk a Brown Korpusszal a mintavétel tárgyalásakor (1.3.1.1.). Feleslegesnek tűnik a többi korpuszra vonatkozó minden apró különbséget felsorolni, hiszen a fontosabbakat már említettük a Brown Korpusz ismertetésekor. Így csak arra hívnánk fel a figyelmet, hogy az egyes korpuszok neve már el is árulja, hogy milyen céllal készültek a korpuszok. A Brown Korpusz klónjai szinte kivétel nélkül mind azzal a céllal készültek, hogy az amerikai angol nyelvvel összehasonlíthassák az angol nyelv különböző változatait: az indiai, ausztrál, új-zélandi és kanadai angolt. Természetesen ezek nem csak az amerikai változattal, hanem egymással is összehasonlíthatók. A klónok közül a kivételek a Freiburg–Brown (FROWN) Korpusz, és a Freiburg–LOB (FLOB) Korpusz, melyek nem a nyelvterületek eltérő nyelvhasználatának összehasonlítása céljából, hanem az időbeli összehasonlítás céljából készültek. Minden nyelv változik, de ez a folyamat lassú, így nehezen érzékelhető mindennapjainkban, talán a generációs nyelvhasználati különbségek kivételével. A FLOB és a FROWN Korpusz a Brownhoz képest „generációs különbséggel", harminc évvel későbbi, azaz 1991-ből és 1992-ből származó szövegeket tartalmaz. A korpusz létrehozását Christian Mair kezdeményezte (http://helmer.aksis.uib.no/icame/flob/). A korpuszok készítésekor igyekeztek minél hűebben megtartani a LOB és a Brown Korpusz szerkezetét és jellegét. Így például a LOB Korpusz által is mintavételezett újságokból merítettek, és a LOB-ban szereplő könyvekkel azonos témájú könyveket választottak ki. A FLOB és a FROWN Korpuszra vonatkozó adatok igen pontosak, még a helyesírási hibák javításait is feltüntetik (13. és 14. táblázat).

A Kategória		***Sajtó: riport (alkategória)***	
A01 (Kód)		*The Independent* (Újság címe)	2021 szó
	Dátum	Cikk címe	
	1991. szeptember 4. 6. o.	Stephen Castle: "Labour Pledges Reversal of NHS Hospital Opt–Outs"	001–032

A Kategória		***Sajtó: riport (alkategória)***	
	1991. szeptember 2. 10. o.	Kevin Hamlin: "Singapore's Voters Give Regime a Shock"	034–099
	1991. szeptember 4. 5. o.	Barrie Clement: "Kinnock Looks to Autumn Poll As TUC Toes the Line"	101–165
	1991. szeptember 2. 10. o.	Andrew Higgins: "Peking Polishes Its Image As Major Arrives" (rövidített)	167–232
	Sajtóhiba:	*manager* javítva: *managers* (008)	
		Boyant javítva: *Buoyant* (104)	

13. táblázat: Példa a FLOB Korpuszban szereplő szövegek adataira (http://helmer.aksis.uib.no/icame/flob/kata.htm alapján)

A Kategória		***Sajtó: riport (alkategória)***	
A01 (Kód)		***San Francisco Examiner (újság címe)***	2007 szó
	Dátum	Cikk címe	
	1992. október 6. A1, 18. o.	'After 35 Straight Veto Victories, Intense Lobbying Fails President with Election Offing.'	001–094
	1992. október 6. A-1, 12. o.	'Clinton Follows Rivals to "Talk TV"'	096–187
	1992. október 6. A-1, 12. o.	'Taunting Disrupts Campaign Stroll'	189–236
	Félreérthető kötőjel:	*high-stakes* (83)	

14. táblázat: Példa a FROWN korpuszban szereplő szövegek adataira (http://khnt.hit.uib.no/icame/manuals/frown/KATA.HTM alapján)

A Brown Korpusz és klónjai megvásárolhatók CD-ROM-on az ICAME, azaz a Modern és Középkori Angol Nyelv Nemzetközi Számítógépes Archívumától, melynek honlapja a következő: http://nora.hd.uib.no/icame.html. A CD-ROM regisztrált felhasználói az interneten keresztül is használhatják a korpuszokat.

3.9.1.2. Könyvkiadók korpuszai

Mint azt már korábban említettük, a Cobuild projekt eredményeképpen a Collins COBUILD English Language Dictionary (J. Sinclair, 1987a) volt az első korpuszelemzés alapján készült, korpuszból eredő példákat használó, nem anyanyelvi beszélők számára készült szótár (lásd a 3.5. részt). Ezek után szinte minden kiadó igyekezett saját korpuszt létrehozni azzal a céllal, hogy az anyanyelvi beszélők nyelvhasználatának pontosabb leírása és a tanulói korpuszok hibáinak elemzése eredményeképpen jobb szótárakat és nyelvkönyveket készíthessenek a nyelvtanulók számára. Nem véletlen tehát, hogy ezek a korpuszok csak az adott kiadónak dolgozó szerzők számára elérhetők. Ezen korpuszok puszta léte is jelzi azonban, hogy ma már egyre kevésbé „illik” csak intuíciók és tapasztalat alapján nyelvkönyvet írni, vagy legalábbis angol nyelvűt nem. Ha szeretnénk minél több külföldivel megismertetni a magyar nyelvet, szükségszerű lesz hasonló elvekre épülő nyelvkönyveket és szótárakat kiadni ahhoz, hogy minél eredményesebb lehessen a nyelvoktatás.

Longman Corpus Network

A nem anyanyelvi korpuszok (3.8.) ismertetésekor már említésre került a Longman Corpus Network (LCN) (Longman Korpusz Hálózat), mely öt részből áll: 1) a már említett 10 millió szavas Longman Learners' Corpus; 2) a 30 milliós Longman/ Lancester English Language Corpus – LLELC (Longman/Lancester Angol Nyelvű Korpusza); 3) a szintén 10 milliós Longman Spoken British Corpus (Longman Beszélt Nyelvi Koprusz), mely a BNC részét képezi; 4) a 100 millió szóból álló Longman Written American English (Longman Írott Nyelvi Amerikai Angol Nyelvi Korpusz); és 5) az 5 milliós Longman Spoken American English (Longman Beszélt Nyelvi Amerikai Angol Nyelvi Korpusz) http://www.longman-elt.com/dictionaries/corpus/lccont.html.

Cambridge Nemzetközi Korpusz – Cambridge International Corpus (CIC)

Az utóbbi körülbelül tíz év munkájának eredményeképpen a Cambridge-i Egyetemi Könyvkiadó több száz milliós korpuszt hozott létre, amelynek tartalmát a következő táblázat ismerteti.

Brit angol

Szavak száma	Korpusz
400 millió	Írott brit angol
17 millió	Beszélt nyelvi brit angol, amely magába foglalja a CANCODE korpuszt, amit a Nottinghami Egyetemmel közösen hoztak létre
20 millió	Írott brit tudományos angol
30 millió	Írott brit üzleti angol

Amerikai angol

Szavak száma	Korpusz
175 millió	Írott amerikai angol
22 millió	Beszélt nyelvi amerikai angol, mely magába foglalja a Cambridge-Cornell észak-amerikai beszélt angol korpuszát, melyet a Cornell Egyetemmel közösen az Amerikai Egyesült Államokban gyűjtöttek
7 millió	Írott amerikai tudományos angol
25 millió	Írott amerikai üzleti angol

Nyelvtanulói angol

Szavak száma	Korpusz
15 millió	Tanulói írott angol (a Cambridge Learner Corpus)
5 millió	Hibakódokkal ellátott írott tanulói korpusz

15. táblázat: A Cambridge Nemzetközi Korpusz http://uk.cambridge.org/elt/corpus/cic.htm alapján

Mint a fenti táblázatból is kitűnik, ez a hatalmas adattár már csak egy név alatt fut (CIC), és az egyes alkorpuszok nevei csak a tartalom azonosítására szolgálnak, nem pedig a korpusz megkülönböztetésére.

A Macmillan Kiadó: World English Corpus – Világ Angol Korpusz

A Macmillan Kiadó kb. 220 milliós korpusza a nevének megfelelően igen sokféle összetevőből áll, melyek a következők: 1) Brit angol; 2) Amerikai angol; 3) Világ angol;[23] 4) nyelvtanulói szövegek; és 5) az angol mint idegen nyelv tanításához használt anyagok. Sajnos az egyes csoportok nagyságáról nem adnak pontos felvilágosítást, így mindössze annyit tudunk, hogy az írott szövegek a korpusz 90%-át teszik ki. A szövegek jellege igen változatos. A korpuszt kizárólag a kiadó használja.

3.9.1.3. Történeti nyelvészeti korpuszok

A történeti nyelvészeti korpuszok nem változnak vagy csak nagyon ritkán, ha esetleg valamilyen új dokumentum kerül a kutatók kezébe. A történeti jellegű korpuszok esetében a helyesírási változatok okozhatnak gondot a keresés során. Jelenleg is számos projekt van folyamatban, amelyeknek az a közös célja, hogy az angol nyelv változását a nyelv fejlődésének valamennyi fázisában elemezze. A következő korpuszok tarthatnak számot érdeklődésre:

- The York–Helsinki Parsed Corpus of Old English Poetry (York–Helsinki Óangol Költészet Korpusza) 71 490 szóból áll, szintaktikailag és morfológiailag elemzett. A korpusz alkalmas többek között olyan kérdések megválaszolására, mint a szórend, szintaktikai, morfológiai, valamint lexikai tulajdonságok. http://www-users.york.ac.uk/~lang18/pcorpus.html
- The York–Toronto–Helsinki Parsed Corpus of Old English Prose (York–Toronto–Helsinki Szintaktikailag Elemzett Óangol Prózai Korpusz), mindössze másfél millió szóból áll. http://www-users.york.ac.uk/~lang22/YcoeHome1.htm
- The Brooklyn–Geneva–Amsterdam–Helsinki Parsed Corpus of Old English (Brooklyn–Geneva–Amsterdam–Helsinki Szintaktikailag Elemzett Óangol Korpusza) 106 210 szó http://www-users.york.ac.uk/~sp20/corpus.html
- The Penn–Helsinki Parsed Corpus of Middle English (Penn–Helsinki Szintaktikailag Elemzett Közép Angol Korpusza), melynek két kiadása is létezik. Közép angol prózai szövegek gyűjteménye, melyet díj ellenében bárki használhat. A második kiadás már CD-ROM-on is megvásárolható. Első kiadás: http://www.ling.upenn.edu/mideng/ppcme2dir/ppcme1.html; második kiadás: http://www.ling.upenn.edu/hist-corpora/
- The Parsed Corpus of Early English Correspondence (A Szintaktikailag Elemzett Korai Angol Levelezés Korpusza), melynek munkálatait a Yorki és a Helsinki Egyetem kutatói végzik. A korpusz kb. 2 millió szóból áll.
- The Penn–Helsinki Parsed Corpus of Early Modern English (Penn–Helsinki Szintaktikailag Elemzett Korai Modern Angol Korpusz), melyen a Pennsylvania

[23] A brit és amerikai angolon kívüli főbb változatok (pl. ausztrál angol) szövegeit tartalmazza, valamint olyan országokból származó szövegeket, ahol az angolt második nyelvként vagy az oktatás nyelveként használják (pl. India).

Egyetemen Anthony Kroch és Beatrice Santorini dolgoznak. CD-ROM-on megvásárolható. http://www.ling.upenn.edu/hist-corpora/

3.10. Az angol nyelvű korpuszok áttekintése

Jóllehet lehetetlen minden létező angol nyelvű korpusz bemutatása, egy korábban készült táblázatot szeretnék megosztani az olvasóval. Ebben a táblázatban a legfontosabb jellemzőket próbáltam megragadni, és ezek alapján csoportosítani a korpuszok egy részét. Egyes korpuszok mérete nem változik, másoké viszont igen gyorsan, ezért a táblázat adatait változatlanul hagytuk. Reméljük, hogy jó áttekintést nyújt a korpuszok sokszínűségéről. A táblázat McEnery és Wilson (1996) valamint Kennedy (1998) könyvében szereplő adatok alapján készült.

	Név	Nyelv	Tartalom	Méret (szószám)	Típus	Címkézett	Elemzett	Hozzáférhetőség/ illetékes
Szöveggyűjtemények	APHB, Az Amerikai Vakok Könyvkiadója		széppróza					nem kutatható
	The Bank of English			418 millió	monitor	igen	Helsinki függőségi nyelvtannal	megbeszélés szerint
	CHILDES Adatbank	nagyrészt amerikai és brit angol, de más nyelvek is	gyermek nyelv- és beszédhibák					igen
	CURIA	ó-ír, hiberno-latin, hiberno-angol						
Dialektus korpuszok	Helsinki Corpus of English Dialects			245 000	beszélt nyelv			megbeszélés szerint
	NITCS (Northern Ireland Transcribed Corpus of Speech			400 000				Oxford Text Archives
Történeti korpuszok	ARCHER, A Representative Corpus of Historical English Registers	amerikai és brit	1650–1990		beszélt és írott	folyamatban		Doulas Biber
	Augustan Prose Sample	brit	1675–1705	80 000	olvasásra szánt			Oxford Text Archives
	Helsinki Diachronic Corpus	angol	800–1710	1 500 000	sokféle zsáner			ICAME
	Helsinki Corpus of Early American English	amerikai angol	késő XVII. és korai XVIII. sz.	500 000				University of Helsinki
	Helsinki Corpus of Older Scots	skót nyelv	1450–1700	600 000				Helsinki
	The Lam Peter Corpus of Early Modern English Tacts		1640–1740	500 000	pamflet irodalom			Lépjen kapcsolatba velük
	The Zurich Corpus of English Newspapers (ZEN)	angol	1660-tól a Times kiadásáig		újság			Lépjen kapcsolatba velük
Többnyelvű korpuszok	The Aarhus Corpus of Contract Law	dán, angol és francia	szerződési jog	1 000 000				Karen Lauridsen
	The Canadian Hansard Corpus	francia–angol	Kanadai parlament	750 000			igen	nyers adatok CD-ROM-on
	The Crater Corpus (ITU Corpus)	francia–angol–spanyol	hír- és távközlési	fejlesztés alatt			fejlesztés alatt	

		Név	Nyelv	Tartalom	Méret (szószám)	Típus	Címkézett	Elemzett	Elérhetőség/ illetékes
Egynyelvű korpuszok	Vegyes csatorna	The Birmingham Corpus	főleg brit		20 000 000	90% írott, 10% beszélt			The Bank of English
		The British National Corpus (BNC)	brit angol		100 000 000	írott és beszélt	1 000 000 szó részletesen elemzett	1 000 000 szó csontváz elemzés	CD-ROM-on
		The International Corpus of English (ICE)	az angolt első vagy legfőbb nyelvként beszélt régiók	hasonló a Brown és a LOB-hoz, de a 60%-a beszélt nyelvi	1 000 000 egyenként				Survey of English Usage
		The Nijmegen Corpus			130 000	főleg írott, valamennyi sportkomentárral			Nijmegen
		The Penn Treebank				főként The Wall Street Journal	igen	igen	részletek az ACL/DCI CD-ROM-on
		The Survey of English Usage (SEU)		szövegek 1953–1987	1 000 000	kb. fele írott és fele beszélt (London/ Lund)			Survey of English Usage
		The Oxford Psicholinguistic Database			99 000				The Oxford Psicholinguistic Database
	Beszélt	A Corpus of English Conversation		A London-Lund Korpusz, az 1970-es években hozzáadott formális beszélt nyelvi angol nélkül					
		The London-Lund Corpus		1960-as és korai 70-es évek, beszélgetések, jogi iratok	500 000				ICAME
		The Polytechnic of Wales Corpus (POW)	gyerekek által beszélt nyelvi angol		61 000			szisztematikus funkcionális nyelvtan	ICAME
		The Lancester/IBP Spoken English Corpus (SEC AND MARSEC)	formális beszélt angol		53 000		igen	csontváz	Contract ICAME
Egynyelvű korpuszok	Írott	The Brown Corpus	amerikai angol 1961-től		1 000 000				ICAME
		The Freiburg Corpus	brit 1991-től	LOB-hoz igazodó	1 000 000				Freiburg
		The Guangzhou Petroleum English Corpus		petrolkémiai	411 612				
		The Hkust Science Computer Corpus		számítástechnikai tankönyvek	1 000 000				Gregory James
		The Kolhapur Corpus of Indian English	indiai angol	1978, LOB-hoz és Brown-hoz igazodó	1 000 000				ICAME
		The Lancester Parsed Corpus			133 000		igen	treebank és csontváz	ICAME
		The Lancester-Leeds Treebank			45 000				Prof. Leech
		The Lancester-Oslo/Bergen Corpus (LOB)	brit angol 1961		1 000 000			igen	ICAME
		The Lancester-Leeds Treebank			45 000		igen	teljes	Prof. Leech
		The Lancester-Oslo/ Bergen Corpus (LOB)	brit angol 1961-től		1 000 000		igen	igen	ICAME

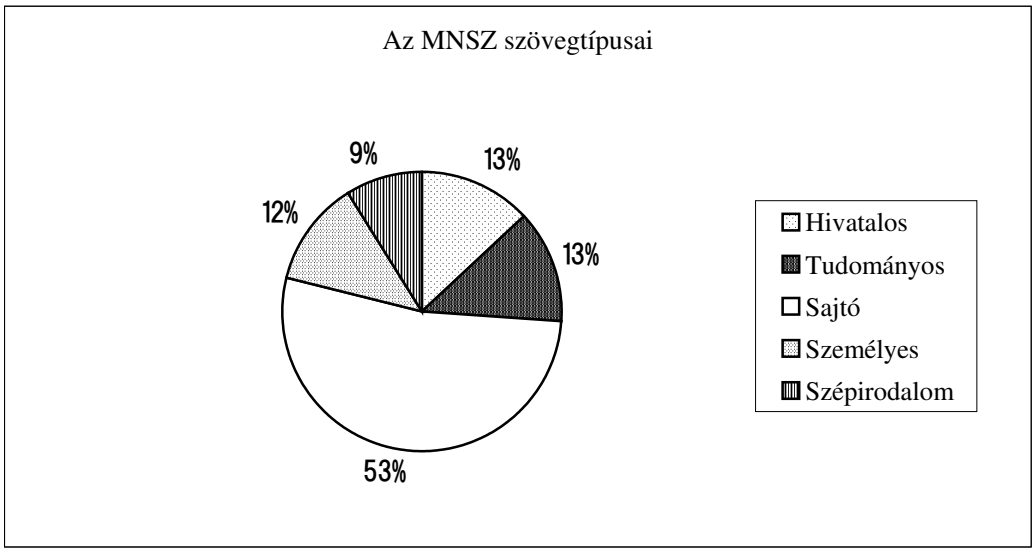
		Név	Nyelv	Tartalom	Méret (szószám)	Típus	Címkézett	Elemzett	Elérhe-tőség/ illetékes
Egynyelvű korpuszok	Írott	The Longman-Lancester Corpus	brit, amerikai, nyugat-afrikai angol		30 000 000				Della Summers
		The Macquaire Corpus	ausztráliai angol		1 000 000				Pam Peter
		The Susanne Corpus		Brown része	128 000		igen	igen, lemmatizált	ICAME
		The Scotish Dramatical Texts Corpus	tradicionális és glasgow-i skót		101 000				John Kirk
		The Tosca Corpus		1976–1986	1 500 000		igen	igen	Nijmagen
		Corpus of English-Canadian Writing	kanadai angol	Brown-ra épült, feminizmus és számítás-technika	3 000 000				

16. táblázat: Angol nyelvű korpuszok összefoglaló táblázata

3.11. Magyar nyelvű korpuszok

3.11.1. A Magyar Nemzeti Szövegtár (MNSZ)

A magyar nyelvű korpuszok tárgyalását természetesen a Magyar Nemzeti Szövegtárral kell kezdenünk, hiszen ez a legátfogóbb, és ezt a korpuszt hozták létre azzal a céllal, hogy „reprezentatívan tartalmazza a mai magyar nyelv jellegzetes megnyilvánulásait” (http://corpus.nytud.hu/mnsz/). A korpusz munkálatait 1998-ban kezdték meg, és jelenleg 153,7 millió szövegszóból áll. A tartalmazott szövegek típusuk szerint a következő százalékban oszlanak meg a szövegszavak száma alapján:



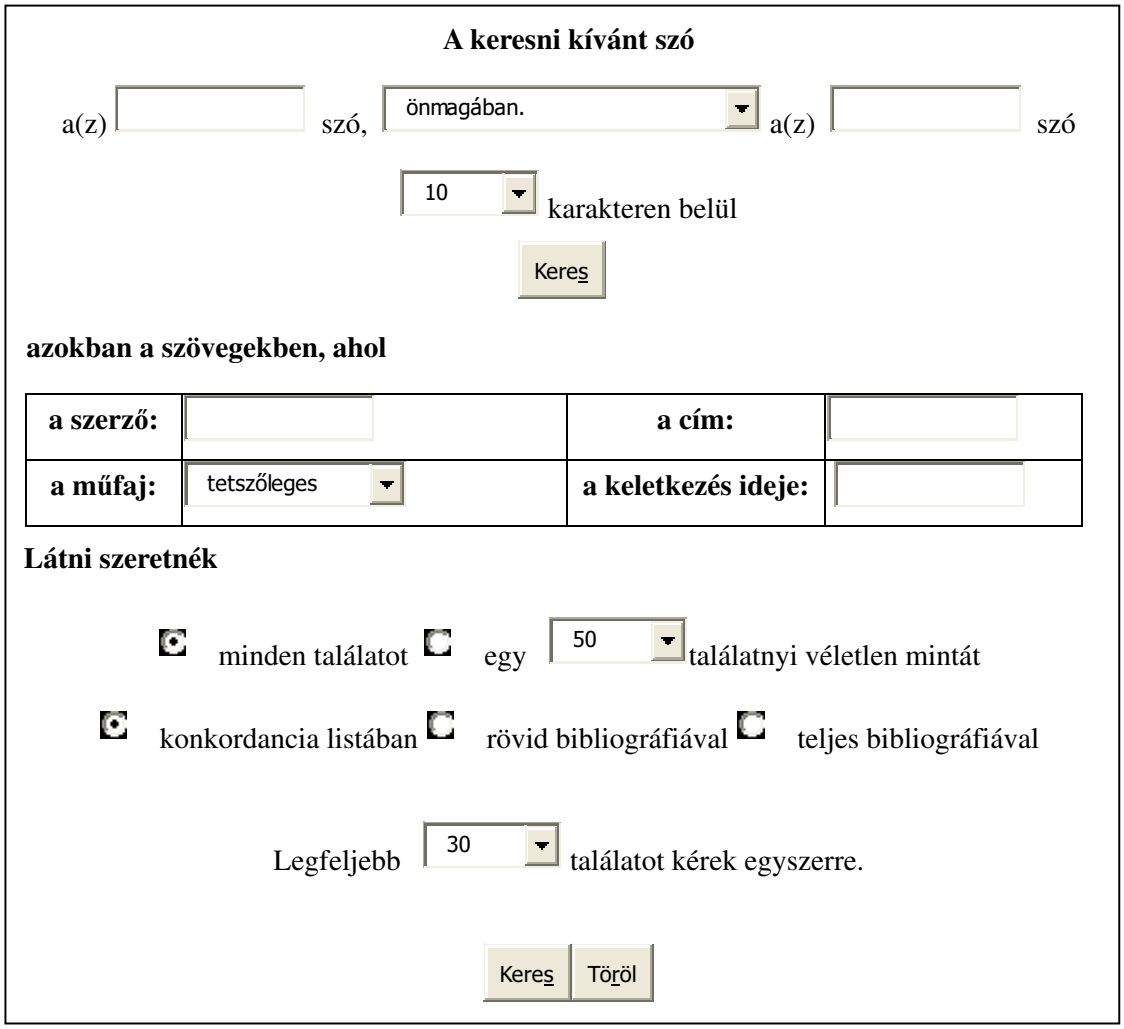
21. ábra: Az MNSZ szövegtípusainak szövegszó szerinti aránya

A korpusz 153 782 228 szövegszóból áll, tehát mérete igen jelentős. Egyértelmű a sajtóból származó szövegek túlsúlya (53%), ami talán egyrészt a könnyű elérhetőséggel is magyarázható. A sajtó szövegeit elektromos formában különösebb nehézségek nélkül „meg lehet szerezni", míg más típusú szövegekhez vagy a jogvédelem, vagy üzleti, politikai okok, esetleg személyiségi jogok miatt nehezebb hozzáférni és a nagyközönség számára hozzáférhetővé tenni. A szépirodalom kategória a Digitális Irodalmi Akadémia szövegeit tartalmazza, ezt folyamatosan bővítik. 40 millió szóra tervezik, és az MNSZ részét fogja képezni. A tudományos szövegek az első fejezetben már említett Magyar Elektronikus Könyvtárból származnak. A hivatalos alkorpusz többek között törvényeket, parlamenti vitákat és szabályokat tartalmaz. A személyes alkorpusz az index.hu internetes fórum szövegeit tartalmazza. Jellegénél fogva, noha írott közléseket tartalmaz, sokszor spontán stílusa miatt az élőbeszéd érzését kelti bennünk. A teljes korpusz anyaga tehát kizárólag írott szövegekből áll, semmilyen beszélt nyelvi alkorpuszt nem foglal magába, ami véleményünk szerint igencsak kívánatos lenne, ha a cél valóban a mai magyar nyelv jellegzetes megnyilvánulásainak tanulmányozása (Váradi, 2002b).

A magyar nyelvet azonban nem csak az országhatárokon belül, hanem azon kívül is sokan beszélik. Nagyon izgalmas és tanulságos lehet mindenki számára a „hazai" használatot összehasonlítani a határokon túli nyelvhasználattal. Erre azonban az MNSZ jelenlegi összetételében nem alkalmas, hiszen jelenleg csak minimális mennyiségben – a szlovákiai Gramma Nyelvi Iroda (http://www.gramma.sk/index.php?lang=hu &lparam=aktualis/korpusz#teteje) szerint 1,5 millió szó határon túlról származó szöveget tartalmaz. Ezeket a szövegek a szlovákiai Új Szó és a romániai Magyar Szó internetes anyagából vették át. A Határon Túli Korpuszt 15 millió szövegszóra tervezik, melyből 6 millió Románia, 4 millió Szlovákia, 3 millió Ukrajna (http://hhrf.org/kmtf/mta/mta_indul.htm) és 2 millió a Vajdaság (http://www.korpusz.org.yu/nyelvikorpusz/sajto_msz_07_2003.htm) területéről kerülne a korpuszba 2005 végéig. Az adatgyűjtés feladata ezekben az országokban azokra a nyelvészekre hárul, akik a Magyar Tudományos Akadémia Nyelvtudományi Intézete által működtetett Nyelvi Irodákban dolgoznak. Az adatfeldolgozás és elemzés nagy részét azonban a budapesti MTA Nyelvtudományi Intézete végzi majd (Kiss, 2004; Bottyán, 2005; Váradi & Oravecz, 1999).

3.11.2. A Magyar Irodalmi és Köznyelv Nagyszótárának korpusza / Magyar Történeti Korpusz

A korpusz 1772 és 2000 közötti szövegeket tartalmaz, melyek összesen 25 millió szövegszót tesznek ki. A XVIII. századból 2 millió szó, a XIX. századból 7 millió, a XX. századból pedig 16 millió szó került a korpuszba. Ez több mint 200 szerző tollából származó 21 000 mű részletét jelenti. A szövegek 82%-a próza, melynek 31%-a szépirodalom. A vers 8,5%-kal, a dráma 5,7%-kal szerepel benne. (Pajzs Júlia, személyes közlés, 2005. március 1.) A korpusz keresőfelületét a http://www.nytud.hu/hhc/ címen találjuk meg. A keresés nem csak a teljes korpuszon végezhető el, hanem bizonyos szempontok szerint általunk választott részein külön is.



22. ábra: A Magyar Irodalmi és Köznyelv Nagyszótárának Korpusza / Magyar Történeti Korpusz kereső felülete

Talán meglepő a következő szó megválasztása, de szándékosan választottam olyan szót, amely nem tipikus az irodalmi szövegek esetében, a hétköznapi életben viszont annál többet hallani. A „hülye” szó begépelésére a következő találatok érkeztek:

1 z folytonos kétség és remény, **hülyeség** okozta rendetlenség, mindenn..
2 korod; s őrült leszesz, vagy **hülye**, / Ha érzelegsz és nem gondolko..
3 m lesz annyi erkölcstelen, / **Hülye**, nyegle, kába. <..
4 Üres kobak nem okos fők, - / **Hülye** bábuk tárgya. <stanz>A..
5 em hevüle, / Néma maradtam és **hülye**. </sectio..
6 korában, tanult meg beszélni, **hülye** volt. 1868-..
7 ekében. Utána Herman Ottó a **hülyék** javára szónokolt. <pa..
8 indulnak az osztályoknak. A **hülyék** osztályában Madách leányár..
9 .. Első : Micsoda **hülyeség**… / Az is bolond, aki a szí..
10 bán! Miért házasodik az ilyen **hülye**? Eredj! - Mi lelt? ..
11 , siketnéma 20, elmebeteg 13, **hülye** 18 ezer. Azóta e..
12 ága; ez az úr néha ír egy-egy **hülye** vezércikket, s azt hiszi, hogy a..

13	szenvednek tőle az utódok. A	**hülyék** és javítóintézetek, a tébo..
14	41 vityillókat épitenek	**hülye** unokáitok! Istenem! ezért dolgo..
15	tettek iránt, gyáva alázatot,	**hülye** megoldást prédikál, kővé mer..
16	fény is éjjé, a gyermekszem a	**hülyeség** odujává, a virág a mérges..
17	yos élet zöld ágát hajhássza,	**hülyének**, csalódottnak, rászedettnek..
18	kolóhoz nem szokott bitang. /	**Hülye** helyett csak " hüllőt " emlege..
19	vasni. Az egész világirodalom	**hülyeség**, Tóth Béla mondja.'' Úgy v..
20	orkij s még vagy ötvenen mind	**hülyék**. Minden irodalom az, - biztatja..
21	ákombákomjait, de így igazán	**hülyeség**. Mindnyájan he..
22	l a báró szigorúan. - Micsoda	**hülye** megjegyzés ez, Bubenyik? ..
23	y odaugatná társainak: - Ti	**hülyék**, mit marjátok egymást? Inkáb..
24	jd akkor mondani, hogy miféle	**hülyék** lehettek azok, kik ezt eltürte..
25	agonizálok. Nélküled egészen	**hülye** vagyok. Egy szegény öreg ember ..
26	k össze, nem pirulván ezzel a	**hülye** váddal alapozni meg a maguk mozg..
27	s valakinek, de minden esetre	**hülye**, aki mindig kitünöen tudta, mit..
28	a Tilala-tóról, a Bodor Ákos	**hülye**, kitalált taváról. A Tilala-to..
29	yos élet zöld ágát hajhássza,	**hülyének**, csalódottnak, rászedettnek..
30	agyok; vagy nekem: akkor ők a	**hülyék**. Mindkét esetben köztem és k..

23. ábra: Találatok konkordancia formában a Magyar Irodalmi és Köznyelv Nagyszótárának korpusza / Magyar Történeti Korpuszban[24]

Vizuálisan is jól látszik, hogy a keresett szó, vastagon szedve a lap közepén jelenik meg, és nem csak az alapszó, hanem annak toldalékos alakjai is megjelennek. A sor elején levő számra kattintva bővebb kontextusban figyelhetjük meg a keresett szót, és az előfordulásának pontos helyéről is felvilágosítást kapunk. A korpuszból származó információ megjelenítésekor az ékezetes betűket nem változtattuk meg, a honalapon való keresés eredeti formájában közöljük.

10 1890-1990 KOMPOLTHY ZSIGMOND: KÉTSZÁZ ÉVES EMLÉK p. 115
A kiadás éve: **1991**
helye: **BUDAPEST**
címe: **MOZGÓ VILÁG**
kiadója: **MOZGÓ VILÁG ALAPÍTVÁNY**
műfaja: **dráma, színmű**

satái, / E papucs stejgerolja a nemzetet. / Színésznek adtam a csizmámat oda, / Hogy ne nyomorogjon mezítelen. Anischlné : Hát mért nem így kezdte rögtön, maga / Drága? Itt van egy puszi és még egy. / - 115. oldal - Még egy. Maga a legjobb szívű német, / Akit láttak e magyar földtekén. / Ezentúl járjon csak mindig papucsban / És tüzelje lábával a nemzetet. 1. élőkép Első asszonyság : Miféle ricsaj! Észtvesztő csődület! Második asszonyság : A színház felé tart a kíváncsi tömeg. Első : Mid adnak ma? Második : Valami Kétszáz éves emlék… Első : Micsoda **hülyeség…** / Az is bolond, aki a színházba belép! 2. élőkép Áruslány : Bort vizet, kolbászkát, / Ropogós perecet / Hogy megéledjék mind, / Aki ma itt beteg. / Tele van a kosár, / Nem adom túl drágán, / Érezzek jól maguk .

24. ábra: Bővebb kontextus és előfordulás helye

[24] A korpuszból vett mintákban helyesírási korrekciót itt és a továbbiakban nem végeztünk, az adatokat az eredeti formában közöljük.

A legelső szám (10) a találat sorszámát jelzi, majd az író életrajzi adatai, neve, a mű címe, és az előfordulás pontos oldalszáma következnek. A további információk a kiadásra vonatkoznak. Meg kell még említeni azonban, hogy az életrajzi adatokat nem minden esetben találjuk meg. Időnként a megjelenés időpontja látható itt is. Még egy fontos tény az is, hogy a kiadás időpontja jelentősen eltérhet a szöveg létrehozásának időpontjától. A fenti példa esetében nem tudhatjuk, hogy e hosszú életű író mikor is írta ezt a művet. Vajon a húszas-harmincas, vagy a nyolcvanas években? A konkordanciákhoz tartozó bibliográfiai adatokat kétféleképpen jeleníthetjük meg, rövid és teljes változatban. A rövid bibliográfia esetében a következőket láthatjuk:

```
1  1844 SÁROSI GYULA: ARADI VÉSZNAPOK
.. ég megemlítnünk: hogy a vízvész folytonos kétség és remény, hülye-
ség okozta rendetlenség, mindennemű embertelen visszaélés, s a leg-
szebb erkölcsi tények gyakorlata között..
2  1859 DÓZSA DÁNIEL: Az éjmadár.
.. att többé nem dalol, / Hajlik korod; s őrült leszesz, vagy hülye, /
Ha érzelegsz és nem gondolkozol. / Sötétben vagy, én a sötétben látok,
/ Ha nem akarsz elveszni, ké..
3  1865 NYULASSY ANTAL: Vasárnap délután.
.. sátok / Korán iskolába, / S nem lesz annyi erkölcstelen, / Hülye,
nyegle, kába. </stanz></page> </text> </section> <section> <head> d>
1900542003 d> <author> ZSI..
4  1865 NYULASSY ANTAL: Mivelődjünk!
.. tykén; / Hig a fejed lágya; / Üres kobak nem okos fők, - / Hülye
bábuk tárgya. </stanz><stanz>Ártatlanság ékesíti / A leány orczáját; /
Virág mezőt, mosolygó szin /..
```

25. ábra: Előfordulás és rövid bibliográfia

Ha a teljes bibliográfiát választjuk, és csak az 1945 utáni kiadványokat vizsgáljuk, ilyen formában jelenik meg az információ:

```
6527  olyan szöveg van, melynek dátuma: 1945-
 700  helyen fordult elő a keresett kifejezés: "hu2lye"

                          1 .. 30 találat

                 További max. 30 találat …   Menü

-------------------------------------------------------------------------

1   1890-1990 KOMPOLTHY ZSIGMOND: KÉTSZÁZ ÉVES EMLÉK p. 115
A kiadás éve: 1991
helye: BUDAPEST
címe: MOZGÓ VILÁG
kiadója: MOZGÓ VILÁG ALAPÍTVÁNY
műfaja: dráma, színmű

satái, / E papucs stejgerolja a nemzetet. / Színésznek adtam a csizmá-
mat oda, / Hogy ne nyomorogjon mezítelen. Anischlné : Hát mért nem így
```

kezdte rögtön, maga / Drága? Itt van egy puszi és még egy. / - 115. oldal - Még egy. Maga a legjobb szívű német, / Akit láttak e magyar földtekén. / Ezentúl járjon csak mindig papucsban / És tüzelje lábával a nemzetet. 1. élőkép Első asszonyság : Miféle ricsaj! Észtvesztő csődület! Második asszonyság : A színház felé tart a kíváncsi tömeg. Első : Mid adnak ma? Második : Valami Kétszáz éves emlék… Első : Micsoda **hülyeség…** / Az is bolond, aki a színházba belép! 2. élőkép Áruslány : Bort vizet, kolbászkát, / Ropogós perecet / Hogy megéledjék mind, / Aki ma itt beteg. / Tele van a kosár, / Nem adom túl drágán, / Érezzek jól maguk / A ma esti drámán! Korhely : Sert ma belém, / Szőkét, / Fülespohárban. / Ittam elég / lőrét / tegnap a bálban. Áruslány <..

---

2 1945 SZÉP ERNŐ: EMBERSZAG p. 31
A kiadás éve: **1945**
helye: **BUDAPEST**
címe: **EMBERSZAG**
kiadója: **KERESZTES**
műfaja: **próza, epika, szépprózа, regény**

y néztek V. igazgató úrra, mintha az ő bűne volna, hogy az oroszok ilyen lassan haladnak, pláne az angolok … Még mindig csak Angers, még mindig csak St.-L, de hány napja már! - Az a baj, hogy az angol katona nem akar meghalni. (Bizonyos asszony fakadt erre a panaszra, nem is először. Mindig az asszonynép a kegyetlenebb.) Ma este aztán leintette végre V. igazgató úr: - Asszonyom, a mult háborúba negyvennégy hónapig voltam kinn, mindig az első vonalban, méltóztasson nekem elhinni, nem volt egyetlen pillanatom se, mikor meg akartam halni. Senki se akar meghalni kérem. Ilyen szemrehányást ne tegyünk könyörgök azoknak a derék tommiknak akik miértünk is küzdenek. Éppen elegen halnak meg, bár föl lehetne támasztani azokat a szép fiatal fiúkat. A hölgy azt mondta erre, kínos kis nevetéssel (**hülyébb** valami nem jutván hirtelen eszébe): - Kérem nem úgy gondoltam. Van egy fajta ember, az aki megtud rögtön mindent: rögtön megtudja, ha valahol finom szivart lehet venni, amilyen már régen nincs a trafikokban. És megtudja, hogy most már Svájc is ad védőlevelet. Hogy tudtátok meg, hol hallottátok? Csuda emberek. Jobb elmenni Svájcba, ha tényleg el kellene menni, mint Svédországba, ott senkivel egy huncut szót se beszélhet az ember. Még B..

26. ábra: Keresés eredménye teljes bibliográfiával

Jelen esetben az a probléma, hogy nem annyira a szövegek számára lenne kíváncsi az ember, hanem inkább a szavak számára. Nyelvészetileg a szavak száma használható adat, a szövegek száma viszont kevésbé. A különböző választható lehetőségek használatakor minden esetben ajánlatos elolvasni a súgót, hogy elkerüljük a „nincs találat” eredményt, amikor tudjuk, hogy a keresett szónak szerepelnie kell a szövegben. Így jártam, amikor egyetlen évszámot, 1991-et írtam keletkezési időpontra.

3.11.3. Szeged Korpusz (http://www.inf.u-szeged.hu/projectdirs/hlt/)

A honlap adatai szerint 1,2 millió szövegszóból áll, és 167 ezer szóalakot tartalmaz. A szövegeket több szempontból, elsősorban morfológiailag elemezték. Az első változat 2000. szeptember 1. és 2002. június 30. között készült, a második változatot egy 200 000 szavas üzleti szövegeket tartalmazó résszel egészítették ki. Jelenleg a részletes szintaktikai elemzéseken dolgoznak, melynek eredményeképpen „a konzorcium az első magyar treebank (ág elemzési adatbank) alapjait kívánja lefektetni. Az adatbázist a későbbiekben részletes szemantikai információval is el kívánják látni a fejlesztők." A korpusz előzetes regisztrációval letölthető, oktatási és kutatási célokra ingyenesen használható.

A korpusz egyenként kb. 200–200 ezer szavas szövegeket tartalmaz öt kategóriában, az adatokat a http://www.inf.u-szeged.hu/projectdirs/hlt/ lapon levő szegedcorpus_1.0.doc-ra kattintva találjuk meg:

- **Tizenévesek fogalmazásai** (8. és 10. osztályos tanulók fogalmazásai két témában).
- **Irodalmi alkotások** (3 regény, de nem teljes művek).
- **Számítástechnikai témájú szövegek** közül az IDG kiadóvállalat Computer-World Számítástechnika c. újságjának összegyűjtött számai, valamint Kis Balázs: *Windows 2000 – haladó könyv haladó szoftverhez* c. könyvének 4., 5., 6. fejezete szerepel a korpuszban. Ez a szöveg az interneten gyakran előforduló számítástechnikai, technológiai jellegű szövegeket reprezentálja.
- **Újságok** (a Magyar Hírlap, a Népszabadság, a Népszava, és a HVG egy-egy teljes száma 1999-ből).
- **Jogi szövegek** (a CD Jogtár: Hatályos magyar jogszabályok CD-ROM-ról).

Fájlnév	Méret					Több-jelentésű szavak aránya	Szavak száma témakö-rönként
	500 mondatos fájlok száma	Fájl	Szavak	Írásjelek	Többértelmű szavak		
10elb.pl	19	6 696 576	104 818	22 329	62 705		
10erv.pl	15	5 837 983	97 786	20 705	52 837		
8oelb.pl	4	1 305 470	20 454	4174	12 279	57,30%	223 058
utas.pl	11	3 858 926	60 202	15 946	33 719		
pfred.pl	13	3 028 319	46 578	14 391	24 284		
1984.pl	14	5 011 830	80 411	17 631	42 965	53,94%	187 191
CWSzt.pl	14	7 129 539	124 043	21 018	57 159		
Win2000.pl	6	3 365 219	57 937	10 888	25 539	45,44%	181 980
np.pl	10	3 794 470	63 966	11 980	31 463		
nv.pl	3	1 332 442	22 630	3900	11 244		
hvg.pl	6	3 345 649	57647	9810	27 990		
mh.pl	6	2 535 555	43 091	7258	20 678	48,78%	187 334
gazdtar.pl	15	7 893 363	134 872	22 908	64 165		
szerz.pl	9	5 174 260	87 314	15 807	42 416	47.97%	222 186
Összesen:	**145**	**60 309 601**	**1 001 749**	**198 745**	**509 443**	**50,86%**	**1 001 749**

17. táblázat: A Szeged Korpusz összefoglaló adatai

3.11.4. Magyar dalszövegek

A magyar dalszövegek egy nagyobb korpusz részét képezik, mely más nyelvek dalszövegeit is tartalmazza. A korpuszt a Lieder and Songs Text Page: http://www.recmusic.org/lieder/ néven találjuk meg. A 77 darab magyar dalszöveg a következő címen található meg: http://www.recmusic.org/lieder/languages.html?LangId=14.

3.11.5. CHILDES Database: http://childes.psy.cmu.edu/ magyar nyelvű korpusza

A CHILDES adatbázis a gyerekek nyelv- és társalgási készségük fejlődésének vizsgálatát teszi lehetővé. Ez nemcsak szövegek tárolását jelenti, hanem az átírt anyagok számítógépes elemzéséhez szükséges programok és egyéb segédprogramok is ingyenesen a kutatók rendelkezésére állnak. Ilyen segédprogram például az átírt anyagokat a digitális videóhoz vagy audióhoz kapcsoló program. Az adatbázis nagy része ugyan angol nyelvű, de 22 más nyelvű adatbázis is szerepel benne, többek között magyar nyelvű is. A magyar nyelvű adatbázis zip formátumban tölthető le.

3.11.6. A Hunglish Korpusz

A Média Oktató és Kutató Központ (MOKK) által készített korpusz magyar és angol szövegek párhuzamos tára, melyet 50 millió szövegszóra terveztek. A honlapon (http://lab.mokk.bme.hu/eszkozok/hunglishkorpusz) található információ szerint a szövegek egyrészt az interneten hozzáférhető forrásokból erednek (EU-jogadatbázis, Gutenberg Projekt és a Magyar Elektronikus Könyvtár, magyar vállalatok üzleti jelentései, és egyebek), másrészt honosítási projektek szövegéből (szabad szoftverek stb). A korpusz pontos méretére vonatkozó adatok is megtalálhatók a honlapon.

Részkorpusz	dokumentumok	szövegszó, 1000	mondatpár, 1000
EU-jogadatbázis	10000	25000	1400 max
Gutenberg-MEK	150	14000	990
Szoftverhonosítás	6 (strange)	600	140
Üzleti jelentések			
Filmfeliratok	400	2500	390

18. tábázat: A Hunglish Korpusz mérete

3.11.7. A Magyar Webkorpusz

A korpuszt 2003-ban a MOKK készítette Szószablya projektjének keretében. A szövegek nem gondos válogatással kerültek a korpuszba, hanem teljesen automatizált szűrés útján az internetről. A korpusz mérete ennek megfelelően hatalmas:

korpusz	oldalak (millió)	szövegszó (millió)	szóalak (millió)
teljes	3,5	1486	19,1
40%	3,125	1310	15,4
8%	1,918	928	10,9
4%	1,221	589	7,2

19. táblázat: A Webkorpusz mérete (http://lab.mokk.bme.hu/eszkozok/webkorpusz/)

Nemcsak maguk az eredeti szövegek, de a belőlük készült gyakorisági szótárak is szabadon letölthetők. A MOKK lapjáról egy nyílt morfológiai programcsomag (hunmorph) is letölthető.

3.12. Egyéb nyelvek korpuszai

Számos olyan szervezet létezik, amely azt a célt szolgálja, hogy széles körben hozzáférhetővé tegye a korpuszokra vonatkozó információkat. Így ezen szervezetek honlapján nem csak egy, hanem sok különböző nyelvre és korpusztípusra vonatkozó információ szerepel. Példaként említenénk az ELRA-t, azaz European Language Resources Association-t (Európai Nyelvi Források Társulása), amelyet 1995 februárjában alapítottak Luxemburgban, és non-profit szervezetként működik (honlapja: http://www.elra.info/). Ugyanakkor alapították a végrehajtó szervezetét is ELDA néven (Evaluation & Language Resources Distribution Agency), amely a hozzáférést ténylegesen biztosítja (http://www.elda.org/rubrique1.html). Sajnos az ilyen jellegű honlapokról „beszerezhető" korpuszok – még tudományos célokra is – szinte kizárólag díj ellenében érhetők el, és az árak általában intézmények és nem egyének pénztárcájához szabottak. A nyelvek igen széles skálája szerepel itt: japán, arab, török, koreai és egyebek a megszokott angol, német és francia mellett. Az ilyen jellegű honlap általában más, hasonló tárházakhoz vezető linkeket is tartalmaz. Ebben az esetben is a Linguistic Data Consortium (USA) http://www.ldc.upenn.edu/ valamint a Bavarian Archive for Speech Signals (DE) http://www.phonetik.uni-muenchen.de/Bas/BasHomeeng.html honlapja érhető el egy kattintással.

Az ELDA-hoz hasonlóan, a Szövegkódolási Kezdeményezés – Text Encoding Initiative (TEI) honlapját is érdemes felkeresni, hiszen itt található a legteljesebb lista azokról a korpuszokról, amelyek a TEI ajánlását követik. Hozzá kell azonban tennünk, hogy sok esetben csalódnunk kellett, mert az adott nyelvnél feltüntetett linket követve csak a korpusz egy töredéke volt valóban olyan nyelvű, amilyennél szerepelt. A honlap címe: http://www.tei-c.org/Applications/index-lang.html.

Az Essexi Egyetem honlapján számos ingyenesen használható, az internetről elérhető korpusz listáját is megtaláljuk, és ezek nem csak az angol nyelvre vonatkoznak (http://clwww.essex.ac.uk/w3c/corpus_ling/content/corpora/list/index2.html#languages).

Az egyéni honlapok között is sokat szenteltek a korpusznyelvészetre vonatkozó információk és a korpuszok gyors elérését szolgáló linkek felsorolására. Mivel az internet állandóan változik, így érdemes több hasonló honlapot is bevenni a kedvencek közé. Jó példa erre, hogy Michael Barlow honlapját, amit éveken keresztül kiindulási pontnak használtam, egyik napról a másikra nem találtam. Az egyéni honlapok közül itt elsősor-

ban azokat ajánlom, ahonnan mások könnyen elérhetők, nem pedig azokat, ahol a tényleges információ megtalálható. A legtöbb honlap angol nyelvű, így mindenkinek ajánlom az alapvető kulcsszavak megtanulását.

Michael Barlow új korpusznyelvészeti lapja: http://www.athel.com/corpus.html. Nem csak angol nyelvű szövegforrásokhoz és korpuszokhoz találunk linkeket, hanem a nyelvek széles skálájához is. Ezen kívül cikkekhez, bibliográfiákhoz, szoftverekhez, és ami még ennél is hasznosabb lehet, német és angol nyelvű korpusznyelvészeti kurzusok internetes változatához juthatunk el erről a lapról.

David Lee saját honlapját (http://www.devoted.to/corpora) sajnos nem értük el, de számos más szerveren megtalálható mása, pl. a Lancasteri Egyetem szerverén: http://lingo.lancs.ac.uk/devotedto/corpora/home/corpus_resources.htm. A linkek a szövegbe vannak ágyazva, így ajánlott annak végigolvasása.

Przemek Kaszubski honlapja (http://www.staff.amu.edu.pl/~przemka/) jól áttekinthető és könnyen kezelhető. Szöveget alig találunk rajta, a linkek magukért beszélnek. A „Corpus Linguistics on the Web”-re kattintva a következő lapról bibliográfiákhoz, korpuszokhoz, letölthető programokhoz, folyóiratokhoz, és más hasznos forrásokhoz vezető linkeket érünk el, szintén nagyon könnyen áttekinthető formában.

Mielőtt az egyes nyelvek korpuszaira térnénk, fontos megjegyeznünk, hogy feltételezzük azt, hogy hozzánk hasonlóan mindenki elsősorban azon korpuszok iránt érdeklődik, amilyen nyelven ért. Ezért nem tartottuk szükségszerűnek, hogy minden nyelv estében mindent aprólékosan magyarra fordítsunk.

3.12.1. Német nyelvű korpuszok

A német nyelvű korpuszok iránt érdeklődőknek jó kiindulási pontot jelenthet a Birminghami Egyetem Német Tanszékén dolgozó Bill Dodd honlapja (http://www.german.bham.ac.uk/dodd/), akinek *Working with German Corpora* (2000) címmel megjelent könyvét is haszonnal forgathatják. A honlapról a Corpus Linguistics linket követve további hasznos információ érhető el. A http://www.german.bham.ac.uk/dodd/verursach.htm címen ízelítőt láthatnak a konkordanciák felhasználási lehetőségeiből is.

3.12.1.1. A negr@ korpusz

A Saarlandi Egyetem Számítógépes Nyelvészeti és Fonetikai Tanszéke hozta létre a 20 602 mondatból, 355 096 szövegszóból álló szintaktikailag elemzett, német újságok szövegéből álló korpuszt. A korpuszt minden egyetem és non-profit kutatóintézet ingyenesen használhatja.

3.12.1.2. A Tiger Korpusz

Körülbelül 700 000 szövegszóból (40 000 mondat) áll, melyek a Frankfurter Rundschau című újságból származnak. A korpuszt félig automatikus módon címkézték és szintaktikailag is elemezték.

A Német Nyelvi Intézet (Institut für Deutsche Sprache, Mannheim IDS) korpuszai:

3.12.1.3. Freiburger Korpus

A korpusz 224 szövegből, 700 000 szövegszóból áll. Nagyrészt 1966 és 1972 között állították össze a Német Nyelvi Intézet (Institut für Deutsche Sprache, IDS) projektjének keretén belül. A projekt célja a beszélt német nyelv nyelvtani és stilisztikai tulajdonságainak az elemzése volt. A felvételeket a televíziós és rádiós adások mellett magán és nyilvános beszédek és beszélgetések révén nyerték. A beszélők egy része nem tudott arról, hogy felveszik beszédüket. A rádiós és televíziós felvételek esetében viszont a felvételkészítés maga a beszédaktus szerves része volt. Függetlenül attól, hogy melyik csoportba tartoztak, arról senki sem tudott, hogy a beszédüket később nyelvészeti vizsgálatok céljára fogják majd használni. A felvételeket kerekasztal beszélgetésekre, interjúkra, beszédekre, riportokra, és elbeszélésekre osztották.

3.12.1.4. Dialogstrukturenkorpus

A 72 szövegből, azaz mintegy 200 000 szövegszóból álló korpuszt a Freiburgi Egyetem Német Tanszékének kutatócsoportja állította össze az IDS-sel együttműködve 1968–1972 és 1974–1977 között abból a célból, hogy a természetesen előforduló beszélgetések szerkezetét tovább vizsgálják. Így tehát kapcsolódik a Freiburger Korpushoz. Elsősorban rádiós és televíziós interjúkat és kerekasztal beszélgetéseket tartalmaz a korpusz.

3.12.1.5. Pfeffer-Korpus

A Kaliforniai Stanford Egyetemen dolgozó A. Pfeffer és W. Lohnes hozta létre az 1960-as évek elején. A korpusz 398 szöveget, összességében 650 000 szövegszót tartalmaz. A felvételeket Németország, Ausztria és Svájc 56 különböző részén készítették, összesen 400 különböző beszélő segítségével. Minden felvétel körülbelül 12 perces, ami körülbelül 1500 szót jelent. A beszélők kiválasztása statisztikai alapon történt, kor, nem és foglalkozás figyelembevételével. 125 témakörbe sorolhatók a szövegek, és csak egyetlen szöveg készült négy beszélő részvételével.

A Német Nyelvi Intézet korpuszai (Freiburger Korpus, Dialogstrukturenkorpus, Pfeffer-Korpus) az intézet által kidolgozott COSMAS lekérdező program segítségével használhatóak, így együttesen körülbelül 1,5 millió szövegszót lehet a gyakoriságuk alapján konkordanciák segítségével vizsgálni.

3.12.1.6. Telefonbeszélgetések (Brons-Albert, 1984)

A 35 szövegből álló korpusz mintegy 44 000 szót tartalmaz. A kutató, Brons-Albert, 10 hónapon keresztül felvette a saját telefonjára érkező hívásokat úgy, hogy a telefonálók nem tudtak a felvételről. Természetesen a felvételek átírásához és publikussá tételéhez később a telefonálók hozzájárulását kérte és beleegyezésüket meg is kapta. A korpuszban minden egyes telefonálóról pontos információk találhatók: koruk, foglalkozásuk és/vagy iskolázottságuk, milyen dialektust beszélnek, valamint a kutatóhoz való viszonyuk is fel van tüntetve. A korpusz nem elektromos formában tárolt.

3.12.2. Francia nyelvű korpuszok

Joggal várhatná most mindenki, hogy a franciaországi korpuszok listájával kezdjük a felsorolást. Sajnos ez nem olyan könnyű, mert igen szegényes a rájuk vonatkozó információ. Francia nyelven eddig csak egyetlen könyvet találtunk, amely a korpusznyelvészetről szól: *Les linguistiques de corpus* (Habert *et al.*, 1997). Habert (1999) tollából született egy francia nyelvű cikk a CLEF projektről is, mely az érdeklődők számára az interneten könnyen elérhető a következő címen: http://www.biomath.jussieu.fr/CLEF/PresentationCorpusClef.pdf. Kanadában viszont több korpusz is már hosszabb ideje hozzáférhető. Azért talán kezdjük mégis az anyaországgal.

3.12.2.1. PAROLE Francia Korpusz
http://www.elda.org/catalogue/en/text/W0020.html

Összesen 20 millió szót tartalmaz, melyből 13 millió a Le Monde napilapból származik. A további rész könyveket, folyóiratokat és egyéb szövegeket tartalmaz. Megvásárolható. A Le Monde napilap teljes korpusza 1987 január 1-jétől kezdődően tartalmazza a cikkeket, így ez a legnagyobb ilyen jellegű francia nyelvű korpusz. Az újság korpusza azonban évfolyamonként külön is megvásárolható: 313 euróba kerül, ha nem ELRA tagok tudományos célra kívánják használni. Mivel azonban a Le Monde az interneten ingyenesen hozzáférhető, a türelmes nyelvtanár vagy tanuló napi letöltésekkel létrehozhatja saját, ingyenes korpuszát.

3.12.2.2. Francia Beszélt Nyelvi Korpusz

A könyvhöz végzett anyaggyűjtés befejezésekor 51 óra társalgási anyagot tartalmazott, melyet Párizsban, Grenoble-ban, Montpellier-ben és Avignonban vettek fel. Az átírás folyamatban van, egyelőre mindössze 5 szöveget lehet az interneten keresztül elérni. A szövegeken kívül a beszélőkre vonatkozó és az elemzéshez szükséges egyéb adatokat is, pl. beszéd szituációja, elérhetővé teszik a későbbiekben.

3.12.2.3. Kanadai Francia Korpusz

A Francia Nyelv Tanácsa (Conseil de la langue française) 1990-ben már szükségét érezte, hogy olyan adatbázisokat hozzanak létre, amelyek segítségével a Kanadában beszélt francia nyelv használatának elemzését és leírását elvégezhetik. Ahhoz is ragaszkodott a Tanács, hogy ez közös erőfeszítés eredményeként jöjjön létre, és a felhasználók nyelvi kutatásaikhoz, vagy a nyelvészeti segédeszközök elkészítésének céljából ehhez ingyen hozzáférhessenek.

Ezt az ajánlást látva a Nyelvpolitikai Titkárság (Secrétariat à la politique linguistique) 1996-ban számos egyetemmel, nyelvésszel, szociológussal, és más jelentős szervezettel konzultált, akik szívükön viselték Québec nyelvi és kulturális fejlődésének a sorsát. 1997–98 során megindult a program állami anyagi támogatása is. 1997 és 2002

márciusa között több mint 1 millió dollárral támogatták a programot. 2002–2003-ban a program folytatásához további 150 000 dollárt kaptak. Ebből is látható, milyen komolyan veszik, milyen fontosnak tartják Kanadában a francia nyelv ottani változatának leírását és elemzését.

A korpusz 5 québeci egyetem 15 alkorpuszát egyesíti. Ezeket a következő címen lehet elérni: http://www.spl.gouv.qc.ca/corpus/index.html. A leírások alapján több adatbázis jellegű egység is szerepel benne, amelyet mi nem neveznénk igazán korpusznak. Az adattárolást végző egyetemek a következők:

Laval: Université Laval,
UdeM: Université de Montréal,
UQAM: Université du Québec à Montréal,
UQAR: Université du Québec à Rimouski,
UdeS: Université de Sherbrooke.

A példa kedvért álljanak itt a Laval Egyetemen található „korpuszok”: 1) Fichier lexical; 2) Index lexicographique québécois; 3) Base de données lexicographiques francophone; és 4) QUÉBÉTEXT (1, 2, 3, 4). Igazából csak az utóbbi tekinthető korpusznak. Ennek alkotóelemei: 1) Irodalmi szövegek 1837–1863; 2) Irodalmi szövegek 1864–1919; 3) Anglicizmusokról szóló szövegek (1826–1930); és 4) Útibeszámolók (1650–1899).

A Laval és Sherbrooke Egyetemen található a Base de données textuelles ChroQué, amely a XIX. század végétől megjelent francia nyelvre vonatkozó québec-i írásokat tartalmazza. A honlapról elindulva az egyes szerzőkről is jó tájékoztatást kapunk. A korpuszok között meglepődve láttunk olyant is, amely a 14–25 éves fiatalok szexről és a szexuális úton terjedő betegségekről megfogalmazott érzéseit és gondolatait tartalmazta (Message d’amour – Szerelmi Üzenet Korpusz, Montréali Québec Egyetem). Számos korpusz tartalmaz olyan írásokat, amelyek a francia nyelv védelmével, az anglicizmusok kritizálásával foglalkozik.

3.12.2.4. Le corpus VALIFLOUI (Variétés Linguistiques du Français en Louisiane) http://languages.louisiana.edu/French/Valifloui.html University of Louisiana at Lafayette

A korpusz 350 óra beszéd átírását tartalmazza, mely 352 adatközlőtől származik. Ezek 74%-a férfi, és 26%-a nő. A francia nyelv Louisianában beszélt változatának elemzését segíti elő, mind térbeli, társadalmi és időbeli változások vizsgálatát lehetővé téve. A korpusz létrehozásának munkálatait 1996-ban kezdték meg, és folyamatosan bővítik az anyagot. Jelenleg a következőket tartalmazza:

	COLLECTION	DATE
1	Collection Barry Ancelet	(1974–1996)
2	Collection „Les conteurs de la Louisiane”	(1982–1983)
3	Collection Geneviève Fabre	(1970)

4	Collection Otis Hébert	(1970)
5	Collection Catherine Jolicoeur	(1980)
6	Collection Alan Lomax	(1943–1935)
7	Collection Sylvie Marchand	(1970)
8	Collection Harry Oster	1950–1960)
9	Collection Helena Putnam	(1991)
10	Collection Ralph Rinzler	(1960)

20. táblázat: A VALIFLOUI Korpusz adatai

A korpuszt írásban benyújtott előzetes kérés alapján helyben lehet megtekinteni.

3.12.2.5. Le Corpus du Théâtre religieux français du Moyen Âge (Középkori Francia Vallásos Színház Korpusza)

Ez a korpusz a Brigham Young University French Medieval Database Project-jének része, melyet a http://www.byu.edu/~hurlbut/fmddp/ címen érhetünk el. A korpusz 231 szöveget tartalmaz és nyelvtörténeti összehasonlítások végzésére alkalmas jellegénél fogva. Tartalma a következő:

- 5 jeux religieux des XIIe et XIIIe siècles;
- 45 miracles et drames religieux du XIVe siècle;
- 181 mystères des XVe et XVIe siècles.

3.12.3. A Szerb Nyelv Korpusza

A szerb nyelv korpuszáról már esett szó a 3.2.1. részben, ahol történeti jelentőségét hangsúlyoztuk, de magáról a korpuszról nem sokat árultunk el. A következőkben fogjuk ezt pótolni. A korpusz kezdőlapját a következő címen érhetjük el: http://www.serbian-corpus.edu.yu/indexie.htm. Sajnos egyelőre a szerb nyelvű leírás még nem készült el, így ma még csak angol nyelven juthatunk információhoz.

A korpusz 11 millió szóból áll, és öt nagy csoportra osztható. Az elsőben található szövegek a XII-től a XVII. századig terjedő időszakból származnak, egy részük olyan szerzőktől, mint például Domentijan, Teodosije, Danilo Érsek, másik részük pedig levelezésből és egyéb okmányokból áll. A szöveg az eredeti helyesírást követi. A második csoportot a XVIII. és XIX. századi írások (pl. Milovan Vidaković, Gerasim Zelić, Joakim Vujić) teszik ki, itt is az eredeti helyesírás szerinti szöveg szerepel. A harmadik csoportban Vuk St. Karadžić összes művei találhatók, számos alcsoportra osztva. A negyedik csoport a XIX. század második felének alkotóinak munkáiból tartalmaz válogatást (pl. Branko Radičević, Marko Miljanov, Petar Petrović-Njegoš, Jovan Jovanović-Zmaj és Đura Jakšić teljes munkássága megtalálható itt). Az utolsó csoportban a kortárs/modern nyelv hat csoportra osztott mintája található: regények és esszék (126 könyv), költészet (215 könyv), napilap (politika), tudományos írások (136 könyv), politikai írások és végül a belgrádi szürrealisták írásai. Ezek összesen kb. 7

millió szót tesznek ki, tehát a korpusz nagyobb része a modern nyelv adatait tartalmazza.

A korpusz egyes részeinek pontos tartalmát könnyen elérhetjük. Íme egy példa a versekre:

Poetry – the list of books and items **CSL**

Szerző	Cím	Szavak száma
Alečković Mira	Poljana	11163
Alečković Mira	Tri proleća	14070
Alić Salih	Lirski dnevnik	6291
Andrić Mirko	U tišini	1553
Antić Miroslav	Ispričano za proleća	3004
Antologija	Antologija makedonske lirike	17544
Antologija	Antologija novije čakavske lirike	10563
Antologija	Antologija posleratne srpske poezije	9100
Antologija	Četrdesetorica	32280
Antologija	Za istinu	7657
Banjević Mirko	Njegošev spomenik	704
Banjević Mirko	Sutjeska	3449
Banjević Mirko	Zemlja na kamenu	11812

21. táblázat: Példa a tartalom felsorolására

A szerző és cím mellett a műben szereplő szavak számát is feltüntetik. A lista mellett azonban a válogatás kritériumait is megtaláljuk.

Egy másik jellemző példa:

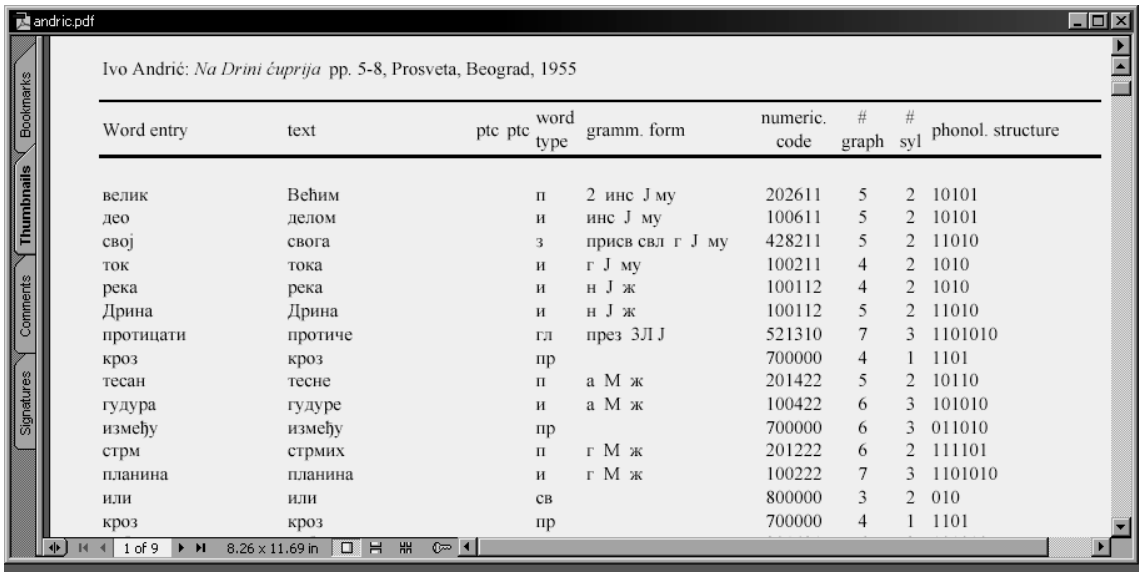
12^{th} – 18^{th} century: the list of books and items **CSL**

Stare srpske povelje i pisma. Knjiga I (I i II deo), Sredio Ljubomir Stojanović, Beograd 1929. *The Old Serbian Charters and Letters. Vol. I (the 1st and the 2nd part),* Collected by Ljubomir Stojanović, Belgrade, 1929.	115 743
Teodosije: Život sv. Save Izdao Đura Daničić, Beograd 1860 (reprint, Beograd 1973). *Teodosije: The Life of St. Sava,* Edited by Đura Daničić, Belgrade 1860 (reprint, 1973).	46 529

Domentijan: Život sv. Simeuna i sv. Save, Izdao Đura Daničić, Beograd 1865. *Domentijan: The Lives of St. Simeun and St. Sava,* Edited by Đura Daničić, Belgrade 1865.	78 305

22. táblázat: A XII-től a XVIII. századig terjedő szövegek korpuszának adatai a szavak számával együtt

Sajnos a korpuszban nem tudunk keresni, de a honlapról elérhetők minták, amelyek pdf fájl formájában letölthetők. Nézzük meg Ivo Andrić egyik művének elemzésének részletét.


andric.pdf

Ivo Andrić: *Na Drini ćuprija* pp. 5-8, Prosveta, Beograd, 1955

Word entry	text	ptc	ptc	word type	gramm. form	numeric. code	# graph	# syl	phonol. structure
велик	Већим			п	2 инс Ј му	202611	5	2	10101
део	делом			и	инс Ј му	100611	5	2	10101
свој	свога			з	присв свл г Ј му	428211	5	2	11010
ток	тока			и	г Ј му	100211	4	2	1010
река	река			и	н Ј ж	100112	4	2	1010
Дрина	Дрина			и	н Ј ж	100112	5	2	11010
протицати	протиче			гл	през 3Л Ј	521310	7	3	1101010
кроз	кроз			пр		700000	4	1	1101
тесан	тесне			п	а М ж	201422	5	2	10110
гудура	гудуре			и	а М ж	100422	6	3	101010
између	између			пр		700000	6	3	011010
стрм	стрмих			п	г М ж	201222	6	2	111101
планина	планина			и	г М ж	100222	7	3	1101010
или	или			св		800000	3	2	010
кроз	кроз			пр		700000	4	1	1101

1 of 9 8.26 x 11.69 in

27. ábra: Ivo Andrić egy művének elemzése

Ha valaki szerb nyelven szeretne információhoz jutni, javasoljuk, hogy vegye fel a kapcsolatot Aleksandar Kostić-csal a következő címen:

Laboratory for Experimental Psychology
Faculty of Philosophy, University of Belgrade
Čika Ljubina 18-20, 11000 Belgrade,
FR Yugoslavia
phone/fax: +381 11 630 542
akostic@f.bg.ac.yu

3.12.4. A horvát nyelv korpusza

A Zágrábi Egyetem Filozófiai Kara Nyelvészeti Intézetének honlapján (http://www.ffzg.hr/zzl/) találunk információt a Horvát Nemzeti Korpuszról. A korpuszépítés munkálatait az 1970-es években kezdték meg. Legutóbbi adataink szerint a horvát korpusz már 53

millió szóból áll. A honlapon nem csak a korpuszra vonatkozó adatokat találtuk meg, hanem a 200 leggyakrabban előforduló szó listáját is. A majdnem 3 milliós irodalmi alkorpuszban, valamint az egyes szerzők alkorpuszában külön is kereshettünk. A korpusz összetételét pontosan ismerhetjük, hiszen a teljes lista mindenki számára elérhető. A fájl neve mellett találjuk a pontos címet, kiadás helyét és idejét. A folyóiratok és a napilapok esetében is pontosan tudni lehet, hogy a fájl mely számokat tartalmazza. Következzen egy példa a korpuszban való keresésre:

30m Corpus of Contemporary Croatian Language (test version) 26.12.2003 05:09:13
Corpus: **30m_test**
Results of query: **škola**

Click the source name to see the wider context ----------
--->
, čiji je član Čekada bio, te Vrhbosanska visoka teološka škola. <p>Skup je otvorio kardinal Vinko Puljić u nazočnos GK9714_62 697 1
bio školovanje u najmanje zahtjevnim programima srednjih škola), a drugi dio učenika bi se u tom devetom razredu za me971217_m01 4190 2
i skupina radova nizozemskih, ali i engleskih slikarskih škola. Treći dio Košine kolekcije čine djela uglavnom hrva VJ981204g 7728 3
športskih natjecanja među učenicima državnih i privatnih škola. Uz dalji poticaj iz predavanja isusovca Carona, Cou GK9631_56 1248 4
kojima su stjecali dragocjena znanja za svoj život. <h3>" Škola mi je dala izvrsno obrazovanje"</h3> <p>Zagrepčanin GK9714_43 2191 5
kolegama glede nekih problema ili pitanja." <h3>"Srednja škola formira čovjeka"</h3> <p>Božo Pavlović, rodom iz Zag GK9714_43 8636 6
je vodila Marina Raspudić te učenici hrvatskih dopunskih škola Stuttgart-Möhringen i Stuttgart-Bad Cannstatt predvo GK9652_58 758 7
životnu istinu da je mnogobrojna obitelj uvijek najbolja škola zajedništva i razumijevanja, ali i odrastanja i samo GK9640_29 12231 8
, a drugi dio učenika bi se u tom devetom razredu za me971217_m01 4190 2
i skupina radova nizozemskih, ali i engleskih slikarskih škola. Treći dio Košine kolekcije čine djela uglavnom hrva VJ981204g 7728 3
športskih natjecanja među učenicima državnih i privatnih škola. Uz dalji poticaj iz predavanja isusovca Carona, Cou GK9631_56 1248 4
kojima su stjecali dragocjena znanja za svoj život. <h3>" Škola mi je dala izvrsno obrazovanje"</h3> <p>Zagrepčanin GK9714_43 2191 5
kolegama glede nekih problema ili pitanja." <h3>"Srednja škola formira čovjeka"</h3> <p>Božo Pavlović, rodom iz Zag GK9714_43 8636 6
je vodila Marina Raspudić te učenici hrvatskih dopunskih škola Stuttgart-Möhringen i Stuttgart-Bad Cannstatt predvo GK9652_58 758 7
životnu istinu da je mnogobrojna obitelj uvijek najbolja škola zajedništva i razumijevanja, ali i odrastanja i samo GK9640_29 12231 8

28. ábra: Keresés eredménye a horvát korpuszban

A keresett szót (*škola*) egymás alatt látjuk, attól jobbra és balra pedig a közvetlen szövegkörnyezetét. A sorok végén levő vonallal aláhúzott betűk és számok kombinációjára kattintva bővebb szövegkörnyezetében figyelhetjük meg a keresett szót.

3.12.5. Szlovén nyelvű korpuszok

3.12.5.1. Szlovén – FIDA

A 100 millió szavas szlovén nyelvű korpusz munkálatai, amelyben a Ljubljanai Egyetem Bölcsészettudományi Kara, a Jožef Stefan Intézet, és két kereskedelmi cég (egy kiadó és egy számítógépes szoftver cég) vettek részt, 1997-ben kezdődtek meg, és 2000 végére fejeződtek be. A kutatást a két kereskedelmi vállalkozás – DZS Általános Kiadó, valamint Amebis szoftver cég – finanszírozta.

Referencia korpuszról van szó, amelynek elsősorban az a célja, hogy lehetővé tegye a szlovén nyelv lehető legszélesebb irányú kutatását. Tehát mind az elméleti, mind pedig az alkalmazott nyelvészeti kutatásokhoz kíván alapot biztosítani. A korpusz mai szlovén szövegekből áll, amelyekben a szlovén szöveg részeként esetenként idegen nyelvű részek is előfordulnak. A szövegek a XX. század második feléből származnak, de érthető módon, a számítógép elterjedése eredményeképpen, jelentős részük az 1990-es évekből származik. A korpusz elsősorban írott szövegekből, vagy előre megírt beszédekből áll. A parlamenti jegyzetek (proceedings) az egyetlen szóbeli része a korpusznak.

A szövegek elsősorban a sajtóból származnak (napilapok, különböző tudományos folyóiratok stb.), de az internetről származó anyagok, valamint beszédek átiratai is kiegészítik a gyűjteményt. A korpusz létrehozása mellett saját kereső programot is kifejlesztettek a kutatók ASP32 néven, mely a korpuszban való keresés webes felületéül szolgál.

A http://www.ijs.si/lit/leposl.html honlapról kiindulva számos irodalmi szöveg elérhető ingyenesen.

A http://bos.zrc-sazu.si/beseda.html címen a Fran Ramovš Szlovén Nyelvi Intézet keresőjével a nemzetközi irodalom szlovén nyelvű fordításaiban is kereshetünk.

3.12.5.2. BESEDA

A BESEDA 112 nagyrészt prózából álló műnek a gyűjteménye, amelyből 98 eredeti mű, 14 pedig fordítás. A korpusz több mint 3 millió szóból áll. Jóllehet a művek 1858 és 1996 között születtek, körülbelül a fele 1962 utánra datálható. A XX. század egyik legjelentősebb írója, Ciril Kosmač teljes életműve is megtalálható, így a szlovén irodalmat szlovén nyelven oktatók számára is igen hasznos forrás lehet. A szövegek annotáltak, és gondosan „meg vannak tisztítva” tipográfiai és egyéb hibáktól. A szlovén korpuszba felvett szövegek eredeti művekre és fordításokra bontott teljes jegyzéke rendelkezésünkre áll.

3.12.6. Cseh nyelvű korpuszok

A Cseh Nemzeti Korpusz Intézetet (Károly Egyetem, Prága) 1994-ben alapították a korpusz megteremtése céljából. A korpusz nem csupán nyelvészek számára, hanem szélesebb kutatói körben is elérhető. A 100 millió szavas korpusz két részből áll: a diakronikus és a szinkronikus vizsgálatokra alkalmas összetevőből. A teljes korpuszból az interneten elérhető változat mindössze 20 millió szó, de a szövegek összetétele és aránya megegyezik a teljes korpuszéval.

A Cseh Nemzeti Korpusz összetétele

<table>
<tr><td colspan="8">A Cseh Nemzeti Korpusz</td></tr>
<tr><td colspan="5">Szinkronikus rész</td><td colspan="3">Diakronikus rész</td></tr>
<tr><td>A ČNKSYN archívum
Az eredeti fájlok</td><td colspan="3">A ČNKSYN bank</td><td>DB
adatbázisok, szótárak</td><td>A ČNKDIA archívum
Az eredeti fájlok</td><td colspan="2">The ČNKDIA bank</td><td>DB
adatbázisok, szótárak</td></tr>
<tr><td rowspan="2"></td><td>SYN2000
100 millió</td><td>ORAL PMK
Prágai beszélt nyelv</td><td>DIAL
tervezett dialektális</td><td rowspan="2"></td><td rowspan="2"></td><td>DIAKORP</td><td>DIAL
tervezett dialektikus</td><td rowspan="2"></td></tr>
<tr><td>PUBLIC
20 millió</td></tr>
</table>

23. táblázat: A Cseh Nemzeti Korpusz

3.12.7. Lengyel korpuszok

A PELCRA (Polish and English Language Corpora for Research and Applications) honlapján (http://www.uni.lodz.pl/pelcra/) találhatjuk a legtöbb információt. A Lengyel Nemzeti Korpusz a BNC mintájára készül. A Lengyel Társalgási Multimédia Korpusz, valamint a Lengyel-Angol Parallel Korpusz munkálatai is folynak. A korpusznyelvészet leginkább a Lodzi Egyetem és Barbara Lewandowska-Tomaszczyk nevéhez fűződik, de megemlíthetjük még Przemek Kaszubski nevét is, akinek honlapját már fentebb megadtuk (http://www.staff.amu.edu.pl/~przemka/).

3.13. Összefoglalás

Ebben a fejezetben először az elektronikus korpuszok előfutárairól esett szó, majd a történetileg jelentős szerepet játszókat vettük számba. Az angol nyelvű korpuszok messze meghaladják az összes többi nyelv korpuszát együttvéve is, amit a fejezet arányai jól tükröznek.

A magyar nyelvű korpuszokat a 3.11. részben mutattam be. Sajnos egyelőre kevés a még magyar nyelvű korpuszokra és korpusznyelvészetre vonatkozó nyomtatásban megjelent szakirodalom, de az utóbbi évek tapasztalatai azt mutatják, hogy napról napra

egyre több információ kerül az internetre. Ennek eredményeképpen nő az érdeklődők száma, és talán egyre többen vállalkoznak majd a korpuszépítésre és elemzésre is. Ha csak az MNSZ példáját nézzük, az elmúlt 2 évben egyre több információ került a honlapra. A felhasználókat a Fórumon feltett kérdéseik megválaszolásával segítik, és a felhasználás során észlelt problémákra is felhívhatják a figyelmet.

A német és francia nyelvű korpuszok száma is viszonylag csekély. A cseh, szerb, horvát, szlovén és lengyel korpuszokra vonatkozó információk is azt bizonyítják, hogy szinte minden országban fontos szerepet játszik a nemzeti korpusz készítése.

Természetesen nem adhattunk teljes képet a korpuszokról, hiszen nap mint nap jelennek meg újabbak, vagy éppen válnak nagyobb egységek részévé. Ez egyben azt is jelenti, hogy mire az olvasóhoz eljut ez a könyv, addigra talán már megszűnik egy honlap, vagy újabb, jelentősebb korpuszokról adnak hírt az interneten. Ha ügyesen választjuk meg a kulcsszavakat, és jó keresővel dolgozunk, akkor szinte bármilyen nyelvű korpuszt megtalálhatunk az internet segítségével. Ha nem járunk sikerrel, akkor se adjuk fel olyan könnyen, hiszen az adott nyelvű, interneten elérhető cikkekből magunk is készíthetünk nyelvészeti vizsgálatok elvégzésére alkalmas korpuszt.

4. A SZOFTVEREKRŐL

4.1. Bevezetés

Napjainkban mindenki a saját bőrén tapasztalhatja, hogy a túlzott információbőség nemhogy segítené a világban való tájékozódást és a világ megértését, hanem inkább elbizonytalanít és esetleg a káosz érzetét is keltheti bennünk. Különösen igaz ez akkor, ha az információ rendezetlen és zuhatagként önt el bennünket. A korpuszok is hatalmas mennyiségű információt tartalmaznak, nemcsak nyelvészeti szempontból. Így tehát az információlekérdezés pontossága, gyorsasága és minősége kulcsszerepet játszik a korpuszok használatával elérhetővé vált információk elemzésében és értelmezésében, ami a használt szoftverek tulajdonságaitól nagymértékben függ.

Az előző fejezetekben többször említettük, hogy az annotált korpuszok jelentős hányadának esetében a korpusz használatához külön szoftvert fejlesztettek ki, amit csak az adott a korpusszal lehet használni. Könnyen belátható, hogy egy bizonyos programot nem lehet két különböző annotációval rendelkező korpusz esetében eredményesen használni, hacsak a program egy részét át nem írják. Ebben a fejezetben olyan programokat igyekszünk bemutatni, amelyek könnyen hozzáférhetőek magánszemélyek számára is. Mivel ezek között magyar nyelvű nem található, azt a megoldást választottuk, hogy az idegen nyelvű program menüit és használatát képekkel illusztrálva részletesen leírjuk. Ezzel szinte egy magyar nyelvű használati utasítást nyújtunk, amit akár más, hasonló program esetében is használhat az olvasó. A legnépszerűbb (és legjobban használható) „fizetős” programok is említésre kerülnek, de elsősorban az internetről ingyen letölthető programokról esik majd szó. Így anyagi lehetőségeitől és nyelvi igényeitől függően választhat az olvasó, hogy melyeket szeretné kipróbálni. A megvásárolandó programok korábbi változatai is sokszor ingyenesen letölthetők, sőt az új verziók bizonyos ideig, általában 2-4 hétig, ha megkötésekkel is, de szabadon használhatók. Javasoljuk, hogy a könyvben szereplő programokat lehetőleg az olvasással egy időben próbálja ki az olvasó.

4.2. A korpuszok készítésekor használt programok

Más eszközökre van szükség és más eszközök használhatók eredményesen, ha a folyóból magunk akarjuk kifogni a halat a vacsorához, vagy ha a boltban élőhalként vesszük, vagy ha félkész mélyhűtött áruként. Így más programokra van szükségünk, ha a használandó korpuszt teljesen magunknak kell elkészíteni, vagy ha „félkész” korpusszal dolgozunk. Még egy hasonlóság a horgászattal az, hogy ha magunk fogjuk a halat, akkor azt esszük, amit a jószerencse a horogra akasztott, és nem válogathatunk túl sokat,

különösen, ha időre kell a vacsorát elkészíteni. A félkész termékek között azonban válogathatunk, és valószínűleg gyorsabb lesz a vásárlás, mint a türelmes pecázás. Végül még egy érv szólhat a félkész termék mellett: sokan irtóznak vagy nem is tudják, hogy hogyan kell halat pucolni. A korpuszkészítést is ajánlott a korpusz használatának jobb megismerése utánra halasztani. Így ebben a fejezetben a korpuszokban rejlő információ lekérdezésére használt programokat ismertetem.

A „nyers" korpuszból a „félkész", fogyasztó számára is megfelelő korpusz elkészítéséhez vezető úton általában a következők történnek – annak ellenére, hogy a korpusz annotációról már az előzőkben szóltunk, szükségesnek érezzük, hogy e fejezet elején dióhéjban felelevenítsük a legfontosabb tudnivalókat. Egészen néhány évvel ezelőttig alapvető feltétel volt, hogy a korpusz csak szövegfájlokból állhatott, amelyek kiterjesztése txt volt. A szövegfájlon belül természetesen nemcsak a szöveg szerepelt, hanem az arra vonatkozó információ is. A szövegre vonatkozó információt a számítógép számára is olvasható módon meg kellett különböztetni a szövegtől. A legegyszerűbb esetben a fájl elején szerepelt a szöveg eredetére vonatkozó információ, és ezen kívül csak a paragrafusokat jelölték. Ezen esetekben a keresés a szöveg egyes elemeire vagy írásjelekre korlátozódott. Ez még meglehetősen „nyers" korpusz, amit viszonylag egyszerűen magunk is létrehozhatunk.

Ha valaki jártas a honlapkészítésben, akkor jól tudja, hogy a honlapon látni kívánt szövegeket kódok veszik közre, amelyek a megjelenítésre vonatkozó információt tartalmazzák, de ezek a honlapon nem jelennek meg. Ehhez hasonló az annotáció jelölése is, ahol ilyen jelek fogják közre a szövegre vonatkozó információt.

A szófaji azonosítás (tagging) esetén minden egyes szövegszót címkével látnak el, és ezeket a címkéket „visszaírják" a szövegbe. Természetesen ezt kézzel is el lehet végezni, ha van rá néhány évtizedünk. Ezt a munkát azonban egy címkéző programmal, azaz taggerrel végzik. A *taggert* előzetes nyelvi elemzések és szólisták (szótárak) eredményei alapján készítik, és még „nem látott" szövegeken tesztelik, hogy a hibaszázalék minél kisebb legyen. Mivel 100%-os pontosságú program nincs, ezért vagy kézileg ellenőrzik, vagy tudomásul veszik, hogy „vannak benne hibák". A szófajilag annotált szövegben már nemcsak szövegszókra kereshetünk, hanem a homográfok (azonos írásképű, de különböző jelentésű szavak) esetében megadhatjuk, hogy milyen szófajt keresünk, például *ég*, főnévként vagy igeként. Vagy kikereshetjük az összes melléknevet anélkül, hogy találgatnunk kellene, vajon például a *tündéri*, és a *mesés* szerepel-e egy szövegben.

A morfológiailag bonyolult nyelvek esetében, mint például a magyar vagy japán, a szófaji elemzésen kívül a morfológiai elemzésre is nagy szükség van. Míg az angolban a lehetőséget külön szóval fejezik ki (*I **can** go home now*.), és így akár egy nyers szövegben is viszonylag könnyen kikereshető, a magyar nyelv esetében (*Hazame**het**ek*.), ha a *-hat/-het* kifejezésre keresnénk, rengeteg más szó is szerepelne a listánkon, például *hetes, heten, hatvan, hetven*, csak hogy néhányat említsünk. A morfológiai elemző programok is már előzetes nyelvi elemzésekre épülnek. A magyar helyesírást és nyelvtant elemző program részét képezi egy morfológiai elemző program, melyet a MorphoLogic nevű magyar cég készített. A morfológiai elemző programokkal nemigen talál-

kozhat önmagában a nagyközönség. A morfológiai elemzés eredménye is „visszakerül” a korpuszba.

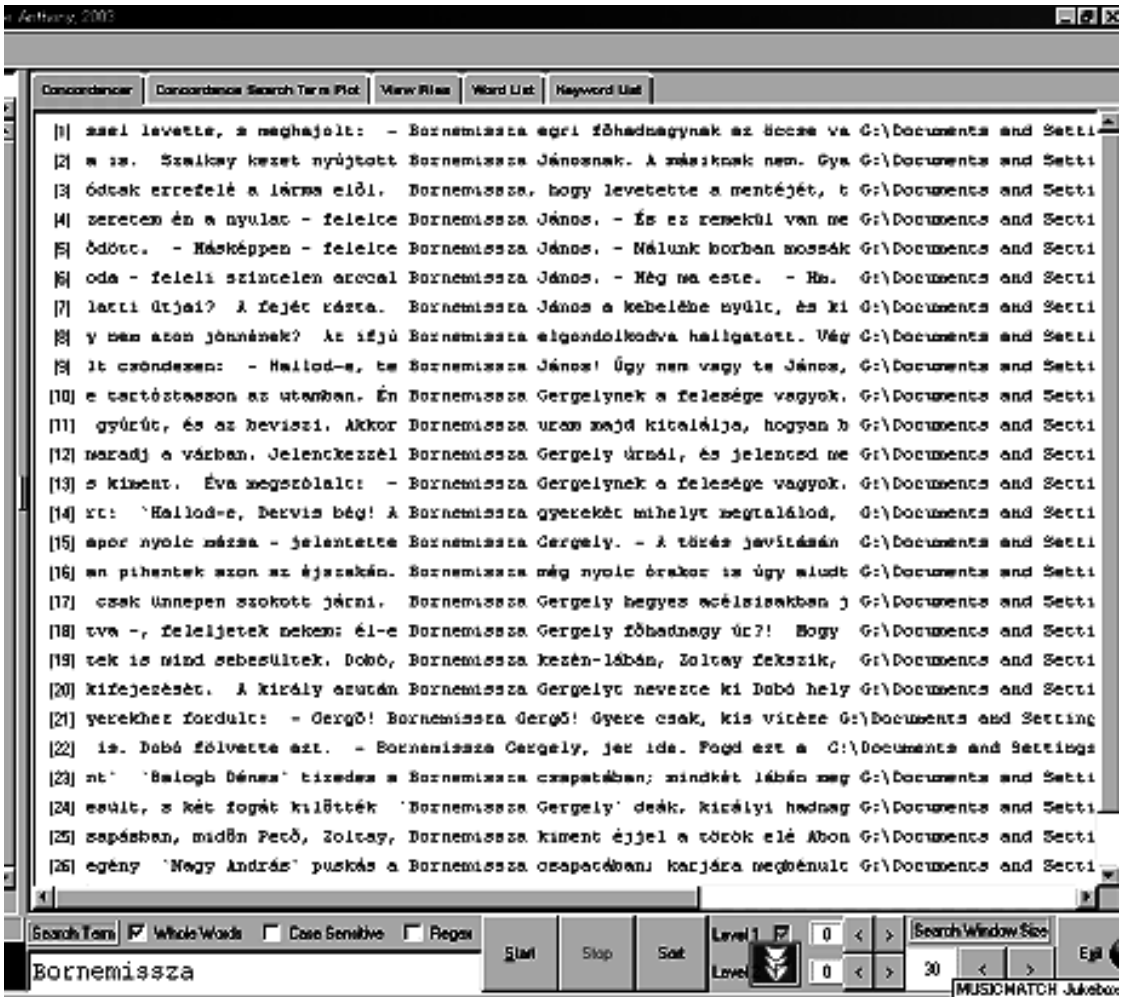
A szintaktikai elemzés az angolban megint csak viszonylag egyszerű, hiszen a szórend viszonylagosan kötött. A kötetlen szórend esetében viszont a morfológia általában segít vagy esetleg egyértelműen meg is határozza a szintaktikai kapcsolatokat. Ezen elemzések eredményei is visszakerülnek a korpuszba. Így tehát a *Tejet ivott lefekvéskor* mondatban szereplő *tejet* szó mellett a korpuszban az az információ is fel lesz tüntetve, hogy ebben a mondatban tárgyként szerepel.

Szó esett már olyan korpuszról is, amely angolul tanuló diákok írásaiból áll. Ebben a korpuszban a nyelvtanár a diákok hibáit és azok fajtáit látta el kódokkal. A kódok lehetővé teszik, hogy bizonyos típusú hibákra keressünk a korpuszban, vagy egyszerűen szám szerint összehasonlítsuk a különböző fajtájú hibákat. Az annotáció variációi szinte korlátlanok, így bárki bármilyen kódot kitalálhat a saját szükségleteinek megfelelően.

Minden olyan információra, amely a korpuszban fel van tüntetve, viszonylag egyszerű programmal is rá lehet keresni. Ebből következik, hogy a már annotált korpusz használata esetén sokkal pontosabb és gyorsabb lesz a keresés és lekérdezés, mint ha csak puszta szövegben keresnénk. Viszont a korpusz előkészítése sok időt, energiát, és ha nem kézzel végezzük, speciális programokat igényel. A következőkben az információt lekérdező programokról esik majd elsősorban szó.

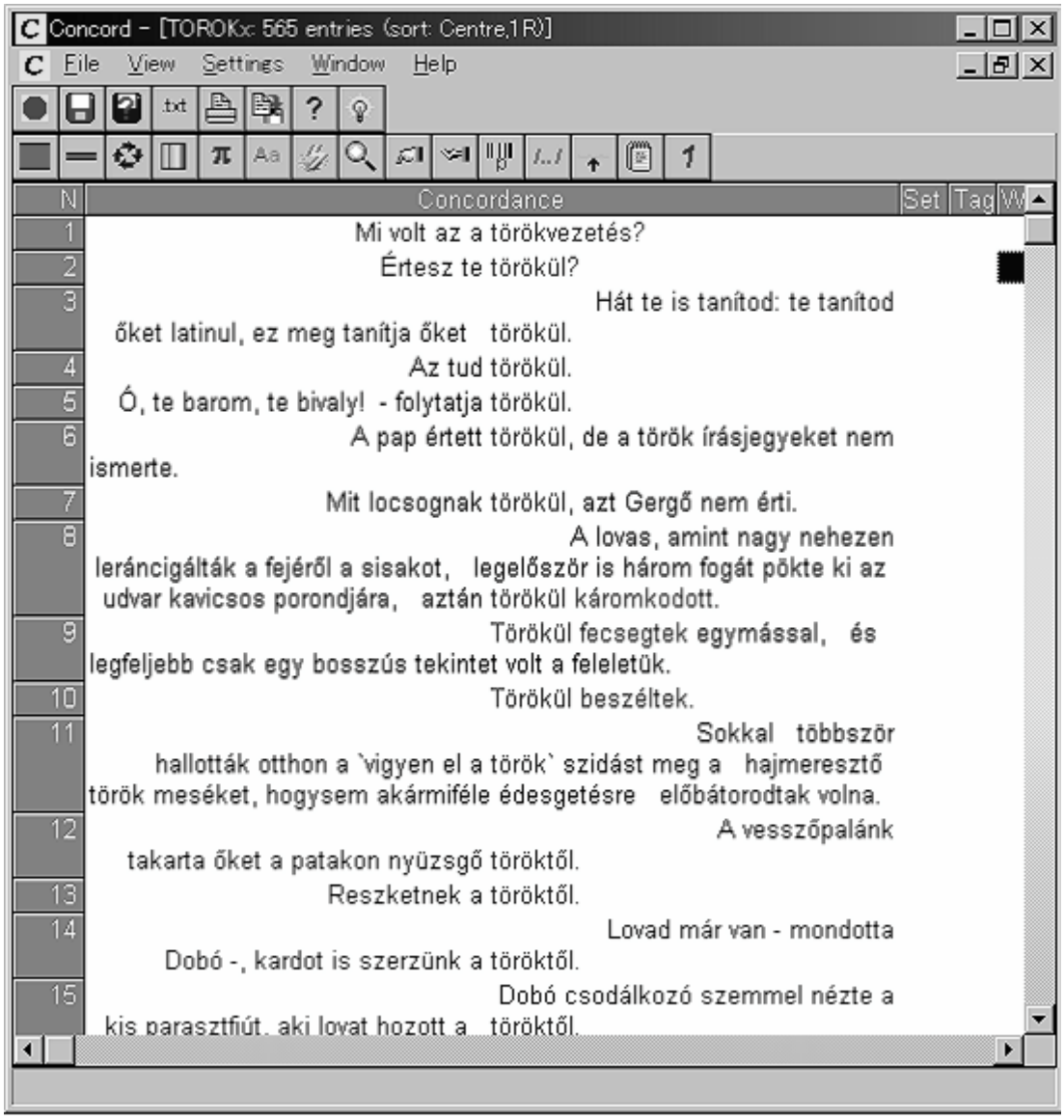
4.3. A konkordanciaprogramok

Bizonyára mindenki, aki használ szövegszerkesztőt, került már olyan helyzetbe, hogy a már elkészített és meglehetősen hosszú szövegben valamit ki kellett javítania, vagy hozzá kellett tennie valamit a már leírtakhoz, de ezt a megfelelő helyen kellett megtennie. Erre valószínű, hogy a Szerkesztés menü Keresés almenüjét használta, melynek segítségével gyorsabban megtalálta a kérdéses pontot. Ez esetben a keresett szó első, második stb. előfordulását könnyen meg is találhatta, de egyszerre csak egy előfordulást lehetett látni. A konkordanciaprogramok abban különböznek ettől a funkciótól, hogy nemcsak kikeresik a keresett elemet, hanem az elem összes előfordulását a szövegkörnyezettel együtt „kimásolják” egy külön ablakba. Így lehetővé válik, hogy egyszerre tekintsük meg a keresett elem előfordulásait a szövegkörnyezettel együtt. A vizuális elrendezés is segíti a gyors felismerést, hiszen a keresett elem mindig a képernyő közepén, azonos helyre kerül. A szövegkörnyezet általában nem teljes mondat, hanem csak annyit mutat egyszerre, amennyi a képernyőre a keresett elem előtt és után kifér. Ez lehet több, mint egy mondat, vagy csak egy mondattöredék. Ezt a megjelenítési formát szokás KWIC, azaz Key Word In Context, magyarul a „kontextusban levő kulcsszó” formának nevezni.



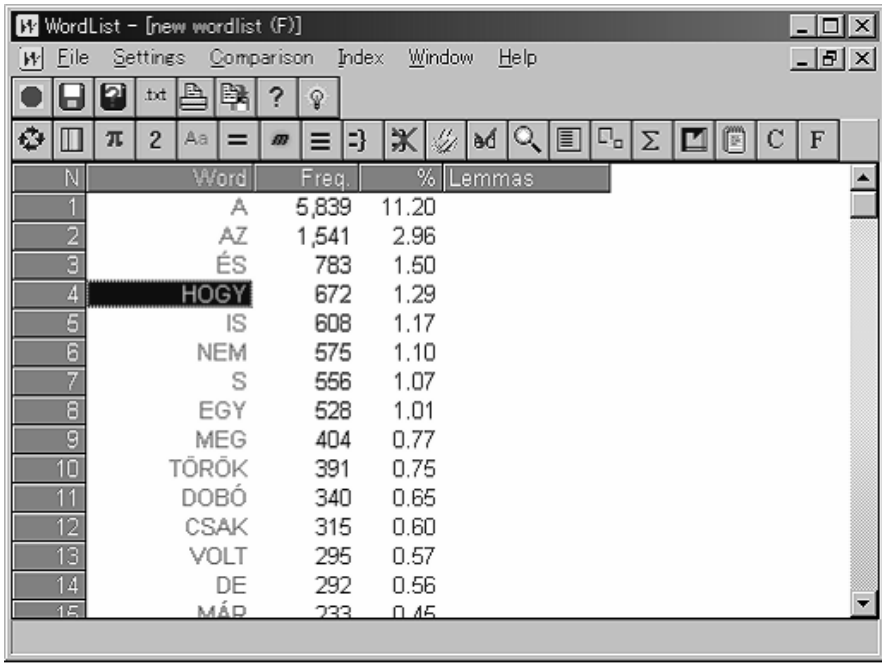
29. ábra: „Bornemissza" konkordanciája az Egri csillagokból (AntConc program)

Ha valakit zavar, hogy nem teljes mondatokat lát a képernyőn, akkor két lehetőség közül választhat. Vagy saját kezűleg törli ki a mondattöredékeket és egészíti ki a hiányzó részeket, vagy olyan programot választ, amely arra is képes, hogy egész mondatokat tegyen láthatóvá a képernyőn. Ebben az esetben azonban előfordulhat, hogy a teljes mondat két vagy több sort foglal el, a mondat hosszától függően. A következő ábra ezt a megjelenítési módot szemlélteti. A 8. és 11. példamondat esetében azt is megfigyelhetjük, hogy ezek három sort foglalnak el.



30. ábra: Konkordanciák mondat formában (*WordSmith* program)

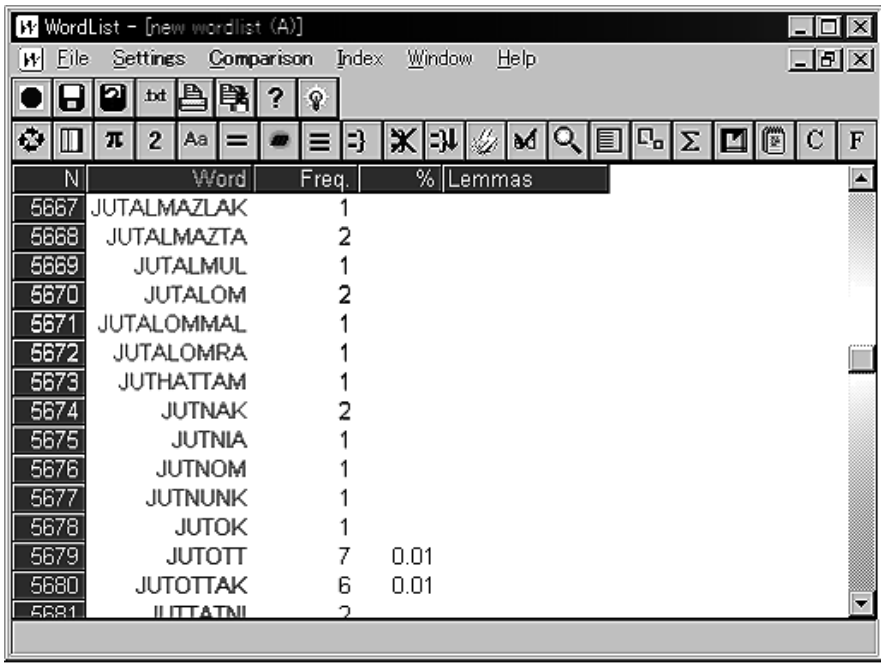
Jóllehet a konkordanciaprogramok erről a funkcióról kapták nevüket, manapság a legtöbb ilyen program számos más funkciót is magában foglal, így nem csupán konkordanciák készítésére alkalmasak, hanem a szövegre és a keresett szóra vagy kifejezésre vonatkozó alapvető statisztikai információkkal is szolgálnak. A szövegszerkesztők, mint például az MS Word, is képesek a teljes szövegben szereplő összes szót megszámolni, de arra már nem képesek, hogy listát is adjanak a szövegben előforduló összes szóról (azaz típusokról, angolul: *types*) és az egyes szavak előfordulásának gyakoriságáról. Az alábbi ábra gyakorisági sorrendben mutatja a szövegben szereplő szavakat. A szó mellett az előfordulások száma és a szöveghez mért százalékos arányuk látható.



N	Word	Freq.	%	Lemmas
1	A	5,839	11.20	
2	AZ	1,541	2.96	
3	ÉS	783	1.50	
4	HOGY	672	1.29	
5	IS	608	1.17	
6	NEM	575	1.10	
7	S	556	1.07	
8	EGY	528	1.01	
9	MEG	404	0.77	
10	TÖRÖK	391	0.75	
11	DOBÓ	340	0.65	
12	CSAK	315	0.60	
13	VOLT	295	0.57	
14	DE	292	0.56	
15	MÁR	233	0.45	

31. ábra: Gyakoriság szerinti szólista az *Egri csillagok*ból (*WordSmith* program)

A szólista minden egyes alakot külön vesz, így ha úgy érezzük, hogy nem külön szóról van szó, ezt összevonással lehet korrigálni. Természetesen ezzel az előfordulási arányok is változni fognak. Az alábbi ábrán a *jutalom* és a *jut* szavak különböző alakjait láthatjuk.



N	Word	Freq.	%	Lemmas
5667	JUTALMAZLAK	1		
5668	JUTALMAZTA	2		
5669	JUTALMUL	1		
5670	JUTALOM	2		
5671	JUTALOMMAL	1		
5672	JUTALOMRA	1		
5673	JUTHATTAM	1		
5674	JUTNAK	2		
5675	JUTNIA	1		
5676	JUTNOM	1		
5677	JUTNUNK	1		
5678	JUTOK	1		
5679	JUTOTT	7	0.01	
5680	JUTOTTAK	6	0.01	
5681	JUTTATNI	2		

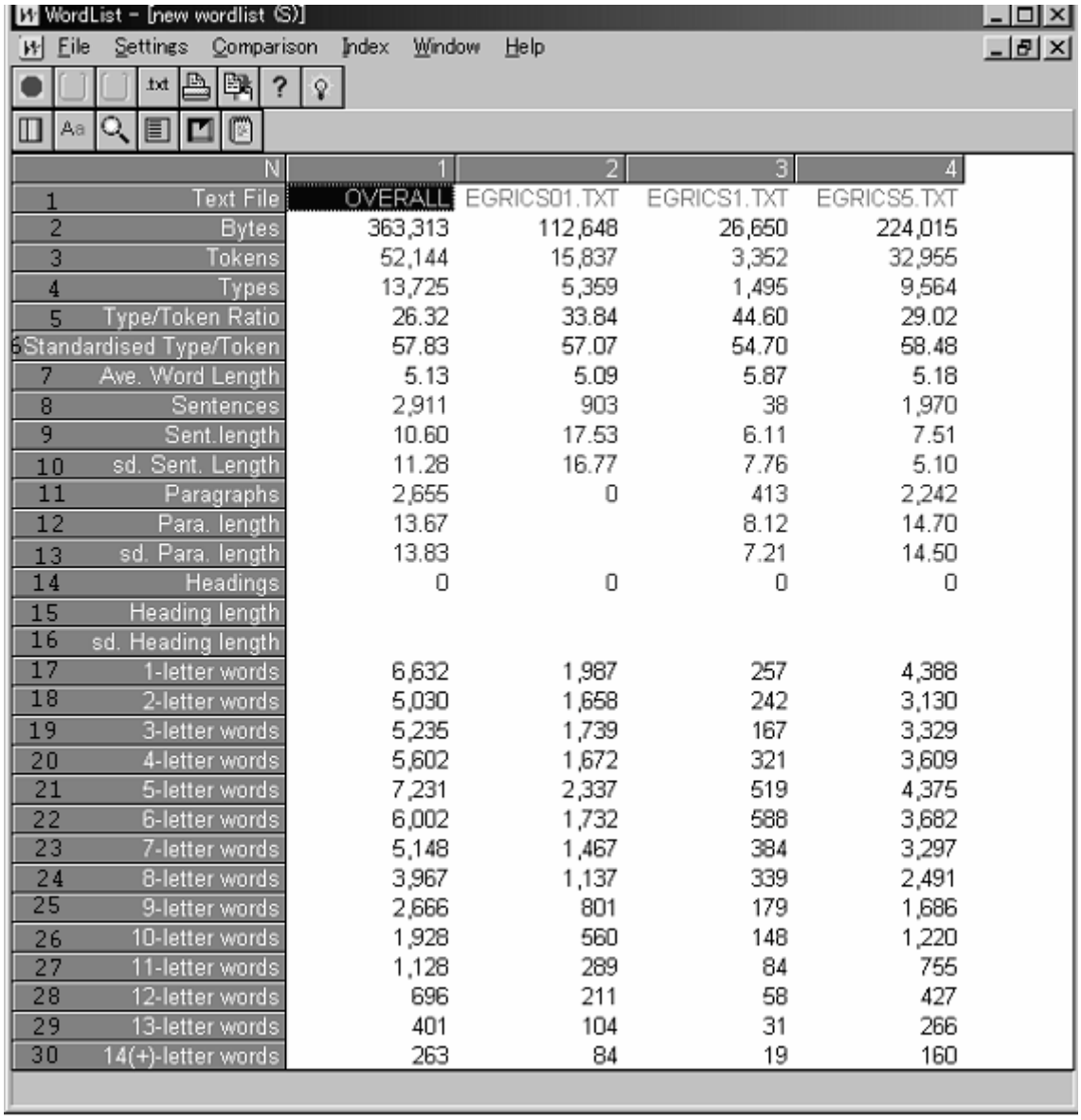
32. ábra: Ábécé szerinti szólista (*WordSmith* program)

A statisztikai adatok programonként változnak. Nézzük meg, egy közkedvelt program, a WordSmith[25] (M. Scott, 1996) milyen adatokkal szolgál. Az alábbi ábra mellett található számok az információ azonosítását segítik.

1. a szövegfájl neve
2. mérete bájtokban
3. szövegszavak száma (összes szó a szövegben)
4. különböző szóalakok
5. különböző szavak és összes szó aránya
6. az 5 pont standardizált változata
7. betűk átlagos száma egy-egy szóban
8. mondatok száma
9. a mondatok átlagos szószáma
10. a 9. pont standardizált változata
11. bekezdések száma
12. bekezdés átlagos hossza
13. 12 pont standardizált változata
14. címsorok
15. címsorok átlagos hossza
16. 15 pont standardizált változata

17–30-ig az 1, 2, 3 stb. betűből álló szavak száma

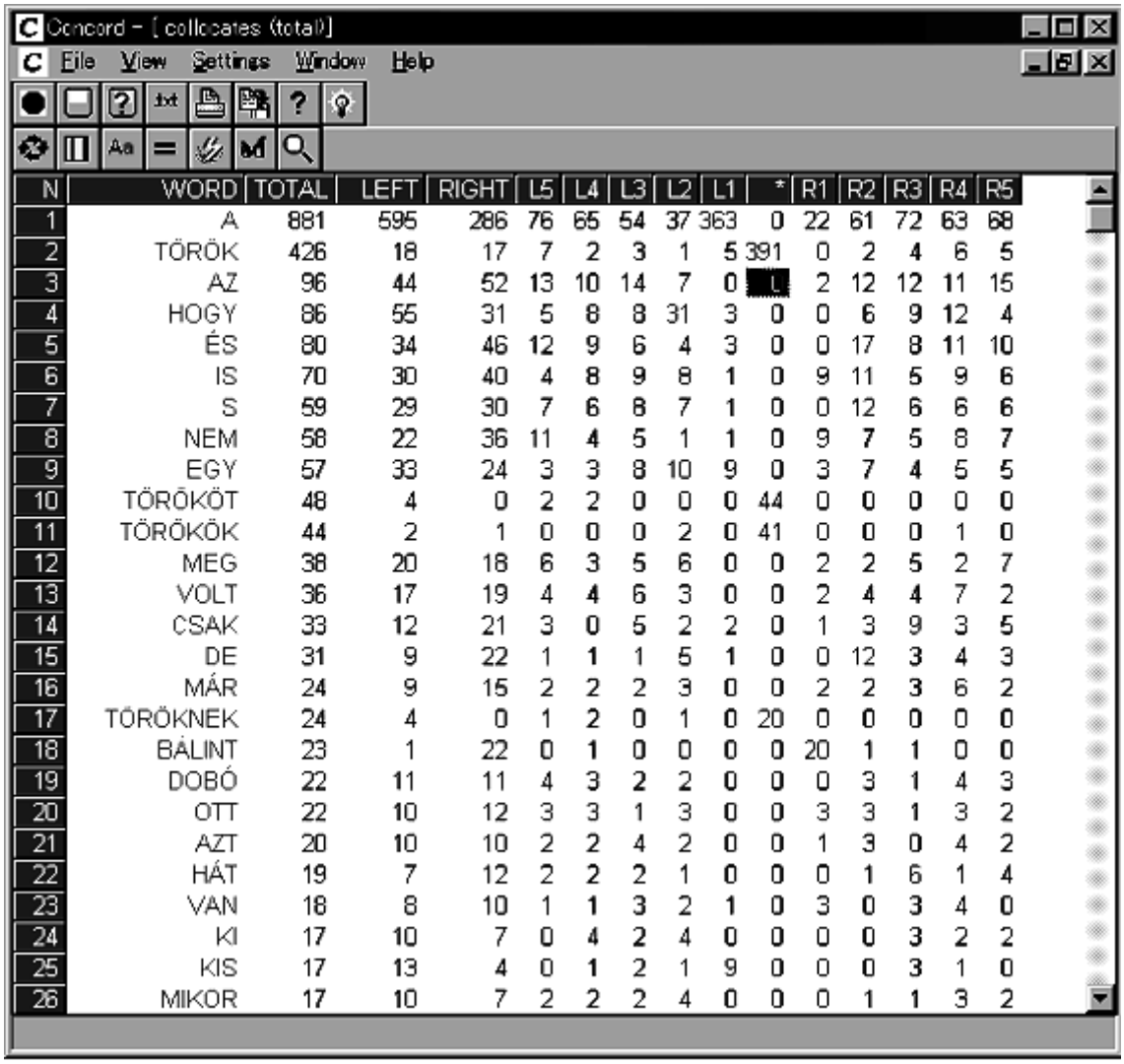
[25] A WordSmith program nagyon népszerű az egyéni kutatók és tanárok körében, jelenleg kb. 50 angol font az egy számítógépen futtatható változata, melyet a készítője honlapjáról http://www.lexically.net/wordsmith/ vagy az Oxford University Press lapjáról http://www.oup.co.uk/isbn/0-19-459400-9 lehet demo változatban letölteni, és a regisztrációs kódot megrendelni. Jelen könyvben kizárólag az ára miatt nem írjuk le részletesen e programot, hanem helyette ingyenesen használhatók kerülnek bemutatásra. A könyv írásakor a program 3. változatát használtuk, a legújabb a 4. változat.



N		1	2	3	4
1	Text File	OVERALL	EGRICS01.TXT	EGRICS1.TXT	EGRICS5.TXT
2	Bytes	363,313	112,648	26,650	224,015
3	Tokens	52,144	15,837	3,352	32,955
4	Types	13,725	5,359	1,495	9,564
5	Type/Token Ratio	26.32	33.84	44.60	29.02
6	Standardised Type/Token	57.83	57.07	54.70	58.48
7	Ave. Word Length	5.13	5.09	5.87	5.18
8	Sentences	2,911	903	38	1,970
9	Sent.length	10.60	17.53	6.11	7.51
10	sd. Sent. Length	11.28	16.77	7.76	5.10
11	Paragraphs	2,655	0	413	2,242
12	Para. length	13.67		8.12	14.70
13	sd. Para. length	13.83		7.21	14.50
14	Headings	0	0	0	0
15	Heading length				
16	sd. Heading length				
17	1-letter words	6,632	1,987	257	4,388
18	2-letter words	5,030	1,658	242	3,130
19	3-letter words	5,235	1,739	167	3,329
20	4-letter words	5,602	1,672	321	3,609
21	5-letter words	7,231	2,337	519	4,375
22	6-letter words	6,002	1,732	588	3,682
23	7-letter words	5,148	1,467	384	3,297
24	8-letter words	3,967	1,137	339	2,491
25	9-letter words	2,666	801	179	1,686
26	10-letter words	1,928	560	148	1,220
27	11-letter words	1,128	289	84	755
28	12-letter words	696	211	58	427
29	13-letter words	401	104	31	266
30	14(+)-letter words	263	84	19	160

33. ábra: Statisztikai adatok az *Egri csillagok* 3 különböző fájljáról

Természetesen arra is kíváncsiak lehetünk, hogy milyen szavak szerepelnek a keresett szóval együtt. Ez nem feltétlenül a közvetlenül mellette levő pozíciót, hanem egy általunk meghatározott „távolságot” jelent. Például a *török** alak keresése esetén a *török, törököt, töröknek*, és egyéb *török* alakkal kezdődő szó is keresett szóként szerepel. Az általunk meghatározott távolság 5 szónyi a keresett szótól balra és jobbra. Arra vagyunk kíváncsiak, hogy milyen szavak szerepelnek leggyakrabban a *török**-kel együtt. A keresés eredménye táblázat formájában így néz ki:


Concord - [collocates (total)]

File View Settings Window Help

N	WORD	TOTAL	LEFT	RIGHT	L5	L4	L3	L2	L1	*	R1	R2	R3	R4	R5
1	A	881	595	286	76	65	54	37	363	0	22	61	72	63	68
2	TÖRÖK	426	18	17	7	2	3	1	5	391	0	2	4	6	5
3	AZ	96	44	52	13	10	14	7	0	[redacted]	2	12	12	11	15
4	HOGY	86	55	31	5	8	8	31	3	0	0	6	9	12	4
5	ÉS	80	34	46	12	9	6	4	3	0	0	17	8	11	10
6	IS	70	30	40	4	8	9	8	1	0	9	11	5	9	6
7	S	59	29	30	7	6	8	7	1	0	0	12	6	6	6
8	NEM	58	22	36	11	4	5	1	1	0	9	7	5	8	7
9	EGY	57	33	24	3	3	8	10	9	0	3	7	4	5	5
10	TÖRÖKÖT	48	4	0	2	2	0	0	0	44	0	0	0	0	0
11	TÖRÖKÖK	44	2	1	0	0	0	2	0	41	0	0	0	1	0
12	MEG	38	20	18	6	3	5	6	0	0	2	2	5	2	7
13	VOLT	36	17	19	4	4	6	3	0	0	2	4	4	7	2
14	CSAK	33	12	21	3	0	5	2	2	0	1	3	9	3	5
15	DE	31	9	22	1	1	1	5	1	0	0	12	3	4	3
16	MÁR	24	9	15	2	2	2	3	0	0	2	2	3	6	2
17	TÖRÖKNEK	24	4	0	1	2	0	1	0	20	0	0	0	0	0
18	BÁLINT	23	1	22	0	1	0	0	0	0	20	1	1	0	0
19	DOBÓ	22	11	11	4	3	2	2	0	0	0	3	1	4	3
20	OTT	22	10	12	3	3	1	3	0	0	3	3	1	3	2
21	AZT	20	10	10	2	2	4	2	0	0	1	3	0	4	2
22	HÁT	19	7	12	2	2	2	1	0	0	0	1	6	1	4
23	VAN	18	8	10	1	1	3	2	1	0	3	0	3	4	0
24	KI	17	10	7	0	4	2	4	0	0	0	0	3	2	2
25	KIS	17	13	4	0	1	2	1	9	0	0	0	3	1	0
26	MIKOR	17	10	7	2	2	2	4	0	0	0	1	1	3	2

34. ábra: A *török szövegkörnyezetében előforduló szavak (*WordSmith*)**

A táblázat első oszlopában szerepel a szó, a második oszlopban az összes előfordulások száma, a harmadik oszlopban a keresett szótól balra való előfordulások száma, a negyedik oszlopban a jobbra való előfordulásoké. Mivel a jobbra és a balra pozíció egytől öt szó távolságig terjed, fontos tudni, hogy milyen közel vagy távol kerülhet ez a szó a keresett szótól. Az ötödiktől a kilencedik oszlopig a balra elfoglalt hely szerinti szám található, a tizedik oszlop a keresett szót jelzi, a tizenegyediktől a tizenötödikig pedig a jobbra levő hely szerinti előfordulás száma látható. Hozzá kell még tennünk, hogy ezek az adatok a mondathatárokat figyelmen kívül hagyják. Az ilyen jellegű információk azonban nagyon fontosak a kollokációk tanulmányozásában. Meg kell azonban jegyeznünk azt is, hogy a *török* szót nemcsak főnévként és melléknévként, hanem a *tör* ige egyes szám első személyű alakjaként is használhatjuk. A 34. ábra eredményei azt sejtetik, hogy itt nem igei értelemben szerepel a *török*, hiszen 363

alkalommal közvetlenül előtte határozott névelő áll. Más esetekben a pontos elemzés érdekében az adott szövegkörnyezet megtekintésével tudjuk csak eldönteni vagy ellenőrizni, hogy főnévi, melléknévi vagy igei értelemben szerepel-e az adott szó.

A konkordanciaprogramoknál olyan funkciót is használhatunk, amely lehetővé teszi, hogy a több szóból álló, de ismétlődő kifejezésekre keressünk. Például az előbb említett *török** milyen két másik szóval alkot kifejezést? A legszembetűnőbb példák ez estben *a török tábor* és változatai, valamint *a török kezére* és változatai voltak. A kereséskor nem adtuk meg, hogy a *török* toldalékmentes alakja esetében az igei jelentést a program figyelembe vegye-e vagy sem, tehát erre az alakra kerestünk. Az eredmény szempontjából ez azonban nem is lényeges, mivel a számokból egyértelműen kiderül, ebben a szövegben a *török* általában jezőként szerepel.

Vannak szavak, amelyeket csak bizonyos szavakkal együtt használhatunk, de azok szinonimájával már nem. Talán sokan emlékeznek még Brachfeld Siegfried paródiájára, amelyben a *dugóhúzó*ból *rejtővonó* lett, hiszen a művész logikája szerint a *dug* és a *rejt*, meg a *húz* és a *von* szinonimák, így akár fel is cserélhetjük őket. (A pontosság kedvéért jegyezzük meg, hogy valójában úgynevezett közeli szinonimák, amelyeknél a felcserélhetőség nem feltétlen kritérium.) Ha idegen nyelven beszélünk, mi is követhetünk el ehhez hasonló hibákat, ha nem a megfelelő szavakat válogatjuk össze, vagy ha nem a megfelelő sorrendben használjuk őket. A magyar nyelvben *fekete-fehér* televízióról beszélünk, amit sok nyelvben ugyanilyen módon, a feketét előre helyezve fejeznek ki, pl. az angolban: *a black and white TV*, franciául: *une télévision en noir et blanc*, a németben: *das Schwarz-Weiß-Fernsehen*. A sok példa ellenére azonban óvakodnunk kell az általánosításoktól, hiszen a japán nyelvben ezt pont fordítva használják: 白黒テレビ (*shiro kuro terebi)*. Ha a példák láttán esetleg valaki arra a következtetésre jutna, hogy a keleti nyelvekben ezt akkor nyilván fordítva mondják, akkor egy kínai példával azonnal óvatosságra intjük. A kínai nyelvben ugyanis, a magyarhoz hasonlóan, a fekete áll elől: 黑白電視 *(heibai dianshi*).

A „kollokáció" néven ismert, a nyelvtanulás és tankönyvírás szempontjából is jelentős fogalom ebben és a következő fejezetben is többször előfordul, így érdemes e fogalmat itt pontosabban meghatározni. Ez talán azért is szükséges, mert fontossága ellenére a magyar nyelvű szakirodalomban alig található meg e kifejezés. A magyar szerzők által készített *Nyelvi fogalmak kisszótárában* (Kugler & Tolcsvai Nagy, 2000) nem találni ilyen szócikket, mint ahogy sem a *Magyar nyelv kézikönyve* (Kiefer, 2003), sem pedig *A nyelv és a nyelvek* (Kenesei, 2004) indexében sem található meg, annak ellenére, hogy a kötetben szerepelnek a szöveggel, gépi szövegfeldolgozással és nyelvtechnológiával foglalkozó írások. Az angol nyelvből fordított *A nyelv enciklopádiájá*ban (Crystal, 2003: 138) azonban már megtaláljuk, hiszen az eredeti műben is szerepel. Az angol nyelvű szakirodalom bővelkedik kollokációkkal foglalkozó könyvekben és cikkekben, és a kutatások eredményeit egyre több kollokációs szótár készítéséhez használják fel.

Kollokáción bizonyos szavak gyakori együttes előfordulását értjük, de ez nem feltétlen jelenti, hogy a kollokánsok (kollokációt képező szavak) közvetlenül egymás mellett állnak, hanem egy bizonyos „távolságon" belül. A szavak természetesen esetlegesen is szerepelhetnek együtt, így joggal merülhet fel a kérdés, hogy hogyan lehet

meghatározni azt a gyakoriságot, amelytől bizonyos szavak együttes előfordulását kollokációnak tekinthetjük. Bonyolult statisztikai képletek és valószínűségszámítási módszerek állnak ehhez rendelkezésre, melyeket szerencsére nem szükséges az olvasónak megtanulni ahhoz, hogy a számítások eredményeit értelmezze. A kollokációk vizsgálatára készült programok már tartalmazzák a számítások elvégzéséhez szükséges kódokat, a felhasználók így azonnal a végeredményt látják.

A kollokáció állandósult kifejezés, de nem idióma, hiszen az idiómák jelentése az alkotóelemek jelentéséből nem áll elő (ez része az idióma definíciójának), pl. a *felkapta a vizet* megértése szempontjából lényegtelen a *felkap* és a *víz* jelentése. Az idiómák általában egy változatban léteznek (nem használjuk azt, hogy **felkapta a tejet* vagy *szörpöt*), így ezeket egy lexikális egységként kezelve könnyen megtanulhatja minden nyelvtanuló. A kollokációkra azonban az jellemző, hogy bizonyos variációs lehetőségek vannak, éppen ezért ezeket sokkal nehezebb a nyelvtanulóknak elsajátítani, mint az idiómákat.

A kollokációk szemléltetésére két melléknevet (*ádáz* és *vad*) vizsgáltunk meg az MNSZ segítségével. A véletlenszerű minta esetében az *ádáz* kollokációiként a következőket találtuk: *csata*, *ellenállás, ellenfél, ellenség, gyűlölködés, harc, küzdelem.* Jóllehet sok esetben az *ádáz* helyett a *vad* melléknevet is használhatjuk ugyan e főnevekkel (*vad gyűlölködés, ellenség* stb.), de ha a *vad* kollokációit is megvizsgáljuk, akkor azt tapasztaljuk, hogy a fenti kifejezések jelentősen gyakrabban fordulnak elő az *ádáz*zal, mint a *vad*dal (pl *ádáz harc* a teljes korpuszban 64-szer fordult elő, de *vad harc* csak 6-szor). A *vad* nagyon sok különböző szóval szerepelt együtt, kevés volt az ismétlődő még az *ádáz*nál 5-ször nagyobb minta esetében is. A többször előfordulók közül említsünk meg néhányat: *dolgok, gyönyör, hullámzás, indulat, kíváncsiság, robbanás, rohanás* és *szenvedély.*

A kollokációk jelentőségét a J. R. Firth vezette londoni iskola ismerte fel elsőként (Firth 1957: 196), és az első kollokációs szótárt is az iskola egyik kiváló képviselője, Harold E. Palmer készítette az 1930-as években (több változatban is). A sok fontos kollokációs szótár közül meg kell említeni Kjellmer (1994) vaskos művét, de Benson & Benson (1993) fontos például az oroszul tanulók számára (és persze az „egzotikus" nyelvek kollokációs szótárait is illene megemlíteni, pl. al-Hafiz 2003-as 373 oldalas arab-angol szótárát – ISBN 9953333793, illetve a kínai Wang Yong és Xie Guofeng 2001-es, 462 oldalas művét – ISBN 7542615084).[26]

4.3.1. A kezdet kezdetén

A kilencvenes évek elején, amikor még kevesen ismerték és használták a konkordanciaprogramokat, számos szótárt kiadó cég is készített egyszerű, de nem igazán olcsó konkordanciaprogramot a kísérletező pedagógusok számára. Abban az időben ez természetesen DOS-ban működő programot jelentett. Nyilvánvaló, hogy a nyelvészek és a nyelvtanárok igényei mások. Így a könnyen kezelhetőség és a világos, könnyen átlátható és szerkeszthető programok lettek népszerűek. A legismertebb programok a következők voltak:

[26] Köszönet jár Cseresnyési Lászlónak az „egzotikus" információért.

- A **Longman Mini-Concordancer** (1989), mely képes volt a szavak számát meghatározni, de viszonylag kis méretű fájlokkal dolgozott (kb. 65 000 szó volt a maximális fájlméret. Talán többen is találkoztak már Chris Tribble és Glyn Jones (1990) könyvével, melynek címe *Concordances in the classroom*, és amely mintegy tanári kézikönyvként szolgált ehhez a programhoz.
- **Micro-OCP** („Micro-OCP", 1988)
- **WordCruncher** (Brigham Young, 1989)
- **Tact**
- **Clan**
- **Free Text Browser**
- **MicroConcord** (M. Scott & Johns, 1993), melynek minimális igénye az MS-DOS 3.0 változata, kb. 200K RAM és 5,25 vagy 3,5 inches hajlékonylemez meghajtó. A program mindössze 156Kb.

Minden fent említett program Windows alapú. A Macintosh programok között azonban már ekkor is megtalálhatóak voltak az ingyenes programok. Mivel Magyarországon a Windows operációs rendszerek sokkal elterjedtebbek, mint a Macintosh rendszerek, a Macintosh rendszereken futó programokat csak érintőlegesen említjük.

Nem hiszem, hogy sok értelme lenne MS-DOS programok leírásával tölteni az időt, hiszen történelmi jelentőségükön kívül semmi gyakorlati haszon nem származik belőle. Senki nem rohanja meg a boltokat, hogy MS-DOS programot vegyen, még akkor sem, ha valóban jól működtek. Az újabb programok olcsóbbak és tetszetősebb felhasználói felülettel rendelkeznek. Nem beszélve arról, hogy a szoftverek készítésekor az eddigi kutatások eredményeit igyekeznek figyelembe venni, és a lehetőségekhez képest azokat úgy alakítani, hogy azok az új kutatási és számítástechnikai igényeknek megfeleljenek. Az igen kedvelt MS-DOS alapú MicroConcord program Windows XP operációs rendszeren már sajnos nem is működik. Az internet jelentőségének megnövekedésével és a „tömegterjesztés" lehetőségével az egyénileg gyártott ingyenes vagy olcsó programok is fellelhetők az interneten.

4.3.2. Internetes felületen futó ingyenes programok

Számos olyan ingyenes program létezik, amely lehetővé teszi vagy az adott programhoz tartozó korpuszban való keresést, vagy pedig a saját számítógépünkön levő fájlban való keresést a program letöltése nélkül. Jó példa erre a Web Concordancer nevű program (http://www.edict.com.hk/concordance/default.htm), mely számos különböző korpuszban való keresés lehetőségét nyújtja. A Bibliától kezdve Drakuláig, a The Times egyes számaitól a „standard" LOB, Brown Korpuszig sok minden megtalálható itt. A keresés eredménye mellett egy szótárhoz vezető kapcsolat is található, amely segít a szavak jelentésének megértésében. Nem véletlen, hogy az angol meghatározás mellett a kínai jelentést is megtaláljuk, hiszen a honlap címéből is kitűnik, hogy Hongkongban került az internetre ez a program. Az alábbi ábra a *house* szó előfordulását mutatja a *The Times* napilap 1995 januárjában megjelent számaiban. A keresett szó 2001-szer szerepel ebben a korpuszban.

Web Concordancer is now searching corpus **TimesJan95.txt** for **house** .....

Concordances for house = 2001	Net Dictionary entries for house

1	't carry a tune from a well to the	**house**	in a bucket" the boys would never
2	acs, general director of the opera	**house**,	said: „A decaying theatre in Cov
3	Scottish businessman, who let the	**house**	as a holiday sporting estate. Mr
4	d their sleeping bags to the White	**House**.	For a man of Robin Renwick's res
5	's Wells Ballet reopened the Opera	**House**	with a performance of The Sleepin
6	s in the committee corridor of the	**House**	have a rough ranking for the Trad
7	, or a combination of these? Tweed	**House**	is a warning to all judges of arc
8	r with people who have purchased a	**house**	with a bit of land and want somet
9	Northern Electric from Trafalgar	**House**	is a challenge to Nigel Lawson's
10	describes a year in his life as	**house**-husband:	a chap who cleans, shops,
11	as been ploughed over and a nearby	**house**,	then a concrete skeleton, has si
12	After a passionate debate the	**House**	approved a constitutional amendment
13	ge, a fishing net loft, the engine	**house**	of a copper mine, a Victorian lau

35. ábra: ***Web Concordancer*** **http://www.edict.com.hk/concordance/default.htm**

A saját szövegeket „feltölthetjük" erre a keresőre, de a magyar nyelvben használt ékezetes betűk miatt nem lesz ideális az eredmény, hiszen ez a program a legtöbb ingyenesen elérhető programhoz hasonlóan, az angolra és esetleg a programozó által beszélt vagy tanult nyelvekre lesz „kihegyezve". Könnyebb olyan ingyenes programot találni az interneten, amely gond nélkül kezeli a japán, kínai vagy koreai írásjeleket, mint a magyarral megbirkózót.

A Brit Nemzeti Korpuszt is hasonló módon használhatjuk a következő címen: http://thetis.bl.uk/lookup.html, természetesen angol szavak kontextusban való megjelenítésére. A Collins Wordbanks Online olyan szolgáltatás, amely lehetővé teszi a Collins Word Web-en rendelkezésre álló korpuszok nyelvi adatainak kutatását. A szolgáltatásért fizetni kell, de a Corpus Concordance Sampler egy 56 millió szavas angol nyelvű korpuszban való ingyenes keresést tesz lehetővé. Ennek használatakor csak 40 konkordanciát láthatunk, de ez is elegendő lehet sok esetben a nyelvtanár vagy tanuló számára (http://www.collins.co.uk/Corpus/CorpusSearch.aspx).

A két legnagyobb magyar nyelvű korpusz, a Magyar Nemzeti Korpusz és a Magyar Irodalmi és Köznyelv Nagyszótárának Korpusza / Magyar Történeti Korpusz is szabadon kereshető az internetes keresőoldalon, de sem a korpuszt nem lehet letölteni, sem pedig saját dokumentumot a keresőben futtatni.

4.4. Konkordanciák készítése

Ebben a részben azt mutatjuk be lépésről lépésre, hogy hogyan és milyen programokkal lehet egy vagy több rendelkezésre álló magyar vagy más nyelvű szöveget konkordanciaprogramok segítségével nyelvi szempontból megvizsgálni. Egyre több olyan program készül, amelynek keresési funkciói és szolgáltatásai megközelítik a régebben csak „profi" intézmények által megfizethetőek szintjét, ugyanakkor már magánemberek

számára is elérhetővé váltak. A viszonylag olcsó és népszerű programok közül a következőket emelnénk ki: Wordsmith Tools 3-as és 4-es változatát (M. Scott, 1999, 2004); a Michael Barlow által készített MonoConc, MonoConc Pro és ParaConc programokat (lásd Barlow, 1999); és a Concordancer (Watt, 2004) nevű programot. Mivel az olvasót most arra kérjük, hogy a fejezet további részét olvasva maga is próbálja ki az itt leírtakat, fontosnak tartottuk, hogy a bemutatásra kerülő programok mindegyike ingyenesen letölthető legyen az internetről. Nagy számuk ellenére kevés azonban az olyan ingyenes konkordanciaprogram, amely alkalmas lenne a magyar nyelvű szövegek vizsgálatára.

Hosszas keresgélés után négy olyan programot választottunk, amelyek különböző igényeket elégíthetnek ki attól függően, hogy milyen célra kívánjuk használni vizsgálataink eredményeit. Máris felhívnánk a figyelmet arra, hogy a pedagógiai alkalmazásokról a következő fejezetben lesz szó, itt csak esetlegesen és röviden utalunk ezekre. A négy program a következő: 1. ConcApp (Greaves, 2003); 2. Simple Concordance Program (SPC) (Reed, 2003); 3. AntConc (Anthony, 2004); és 4. Multi-Lingual Corpus Toolkit (MLCT) (Piao, 2002). Mindegyik programot magyar Windows XP operációs rendszeren futtattuk, és probléma nélkül működtek. Eredetileg azonban nem feltétlenül erre a platformra készültek, de XP környezetben is működnek. Néhány esetben a korábbi változat(ok), mint pl. Win 98 vagy Win 2000-re írottak most is letölthetők. A következő táblázat a legfontosabb, letöltéshez szükséges információkat tartalmazza. Ha elegendő hellyel rendelkezünk a számítógépen, érdemes mindegyiket letölteni és kipróbálni. A http://lingo.lancs.ac.uk/devotedto/corpora/software.htm honlapról mindegyik elérhető a Free Concordancer címszó alatt.

Program neve	**SCP**	**AntConc**	**ConcApp**	**MLCT**
zip fájl mérete	10,8Mb	2,69Mb	teljes 2,55Mb	474kb
utolsó változat	2003	2004	2003	2002
programozó	Alan Reed	Laurence Anthony	Chris Greaves	Scott Songlin Piao
honlap	http://www.textworld.com/	http://www.antlab.sci.waseda. ac.jp/	http://www.edict.com.hk/pub/concapp/	http://www.lancs.ac.uk/staff/piaosl/research/download/download.htm
e-mail cím	A.Reed@textworld.com OR A.Reed@talk21.com	anthony@waseda.jp	ecchgr@hotmail.com	s.piao@lancaster.ac.uk

24. táblázat: Ingyenesen letölthető programok

Minden zip kiterjesztésű fájlt a *winzip* nevű program, vagy más zip kiterjesztésű fájl kezelésére alkalmas programmal kicsomagolunk egy általunk választott nevű könyvtárba. Az SCP program azonnal telepíthető változatban is letölthető. Meglehetősen nagy fájl, így letöltése hosszabb időt vesz igénybe még kábeles internetes kapcsolat esetén is, kb. 20 percbe telt letöltése ebben az esetben. Az AntConc programot letöltése után

azonnal lehet használni, semmilyen telepítésre nincs szükség. A ConcApp programot a setup programmal installáljuk, ezután már a szokásos módon a Start, majd Programok menüpontra kattintva választhatjuk ki és indíthatjuk.

Az MLCT program esetében a program működéséhez szükséges, hogy számítógépünkön legyen egy Java Runtime Environment (JRE) nevű program, amelynek letöltéséhez egy linket találunk a honlapon. A zip fájl kibontása után a mappában levő run_mlct_concordance_jar.bat fájlra való kattintással lehet a felhasználói felületet elindítani. Az első kinyíló ablak azonban egy fekete DOS ablak, és csak egy kis idő elteltével fog megjelenni a második, a tényleges program ablak.

A programok elindítását megkönnyíti, ha a telepítés során létrehozott ikonokat keressük, mert ezekkel lehet a programokat elindítani. A konkordanciaprogramokra általában jellemző, hogy .txt kiterjesztésű fájlokkal működnek. Az újabbak XML, vagy a program súgójában ismertetett egyéb fájl formátummal is működhetnek. A programok tanulmányozása során a Magyar Elektronikus Könyvtárból letöltött *Egri csillagok* szövegfájljaival dolgoztunk. Ajánlott a kisebb méretű fájlokon való kipróbálás, hiszen így gyorsabban meggyőződhetünk arról, hogy egy-egy utasítás kiadása után milyen eredményt kapunk. Ezért sokszor az öt fájlból álló teljes szöveg helyett csak egy szövegfájlt használtunk.

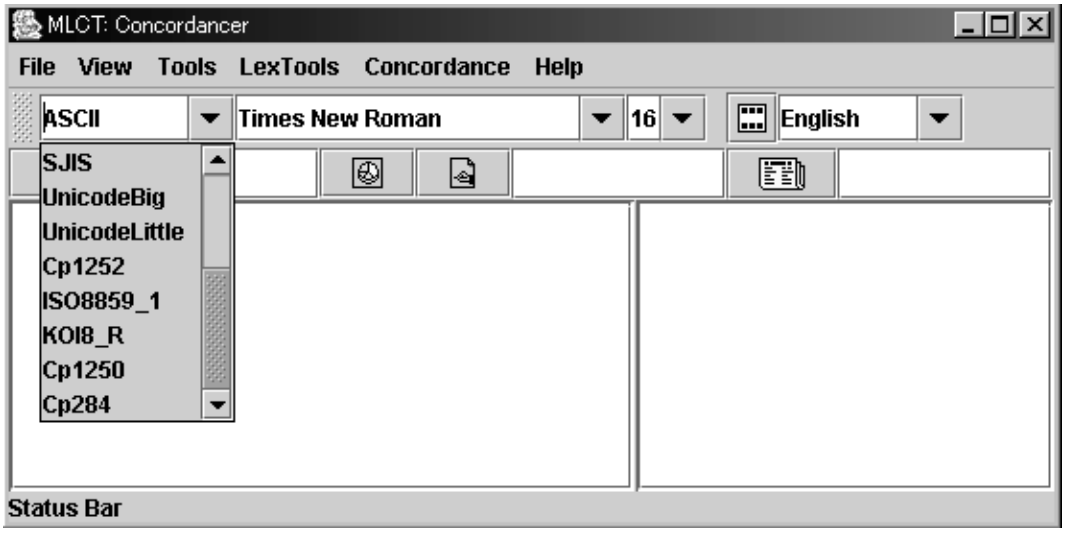
Az itt felsorolt programok mindegyike folyamatos fejlesztés alatt áll, így előfordulhat, hogy ha ma még nem is tudunk valamilyen elemzést elvégezni a programmal, a következő hónapban viszont már igen. Ezért javasoljuk, hogy az olvasó időnként „nézzen vissza" a program honlapjára és töltse le a legújabb változatot. Még egy általánosnak nevezhető dologra kell felhívnunk a figyelmet. A konkordanciaprogramok készítése során egyre pontosabb keresési módokat építenek a programokba készítőik. Itt a regular expression, röviden regex vagy regexp használatára gondolunk. A keresés módjánál sokszor választható, hogy szavakat keresünk-e vagy regexeket. Egy szót is be lehet gépelni regexként. Ezt használtam ki, amikor egy program nem tudta értelmezni a magyar ékezetes betűket. Ha szavakat kerestem, az ékezetes betűket teljesen figyelmen kívül hagyta a program a keresésénél, de ha regexként írtam be az ékezetes szót, akkor „megtalálta". Elégedjünk itt meg azzal a meghatározással, hogy a regex olyan szimbólumok és szintaktikai elemek készlete, amelyekkel szövegszerkezeteket (pattern) azonosíthatunk. Példaként említsük a két leggyakrabban használtat: bizonyára sokan használták már kereséskor a ?[27] vagy a *[28] szimbólumot, de említhetnénk a *keres és cserél* funkciót is az MS Word használatakor. Nyilvánvaló, hogy ezek segítségével pontosabb kereséseket végezhetünk a szövegben. Itt nem foglalkozunk ismertetésükkel, de érdemesnek tartjuk egyéni tanulmányozásukat.

[27] A ? egy karakter, azaz betű vagy szám helyettesítésére alkalmas jel. Például, ha a *t?r* kifejezésre keresünk, a ? helyén állhat bármilyen betű, így a *tar*, *tár, tér*, *tör, tőr, túr, tűr* alakok mindegyikét egyszerre megtaláljuk.

[28] A * tetszőleges számú karaktert helyettesít. Például a *török** keresés eredményeként megtaláljuk a *török, törökül, törökök, töröknek* stb. alakokat.

4.4.1. Az MLCT

Kis mérete ellenére talán a legmélyrehatóbb lexikai vizsgálatokat az MLCT programmal végezhetjük. Ez a program része egy programkészletnek, amely többnyelvű szövegfeldolgozásra és nyelvi vizsgálatok céljára készült. A program indítása után két dologra kell figyelni. A program által használt kódolást, a beállított ASCII-ról Cp1250 vagy Cp1252-re át kell állítani. Ha ezt elmulasztjuk még a fájl megnyitása előtt, az ékezetes betűk nem fognak helyesen megjelenni a képernyőn. A program kétnyelvű dokumentumok vizsgálatára is különösen alkalmas, hiszen két ablakkal működik, és kívánság szerint a jobb vagy a bal oldalon nyithatjuk meg a dokumentumokat. Ha egy nyelvvel vagy fájllal dolgozunk, akkor a bal oldalon nyissuk meg a fájlt, mert az eredményeket mutató ablak a jobb oldali. Az alábbi ábra mutatja a kezdő ablakot és az ASCII megváltoztatására szolgáló menüt. A kódolásokból látható, hogy japán, koreai és kínai szövegeket is vizsgálhatunk a programmal, ha a számítógépünkre a megfelelő kiegészítő programok telepítve vannak, amelyek az ezeken a nyelveken történő szövegszerkesztéshez is elengedhetetlenek. Az ábrán a Times New Roman betűtípus neve olvasható. A jobb oldalon levő nyílra kattintva a legördülő lehetőségek közül kiválaszthatjuk az általunk kedvelt és a vizsgált nyelvnek legjobban megfelelő betűtípust. A betűk mérete (az ábrán 16 pont) hasonló módon állítható a következő legördülő menü segítségével.



36. ábra: Az *MLCT* program kezdő ablaka

A fenti ábrán látható, hogy a nyelv jelenleg angolra van állítva, így az angol nyelv szerinti ábécésorrendet és betűkészletet használja a program. A választási lehetőségek: angol, kínai, koreai, finn és egyéb nyelvek. Az ékezetes betűk miatt soha nem fogunk a programtól hibátlan magyar ábécé szerinti listát kapni, de ezt más programban könnyen korrigálhatjuk. A program a kis- és nagybetűket megkülönbözteti. Az English felirattól balra eső, a dobókocka hatos számát mutató ikonra való kattintással a bal ablakban levő

szöveget automatikusan mondatokra és paragrafusokra oszthatjuk, aminek eredményét a jobb oldali ablakban láthatjuk majd. Az ablak alsó részén, ahol a fenti ábrán jelenleg a Status Bar olvasható, a program a műveletek végzése közben az adott műveletet kiírja, majd a művelet végrehajtása után ismét a Status Bar jelenik meg. Egyedül ez alapján tudjuk csak megállapítani, hogy a program még dolgozik-e az adott műveleten vagy már befejezte azt.

A nyitó ablak áttekintése után nézzük meg a menükben szereplő utasításokat. A File (fájl) menü a következőkből áll:

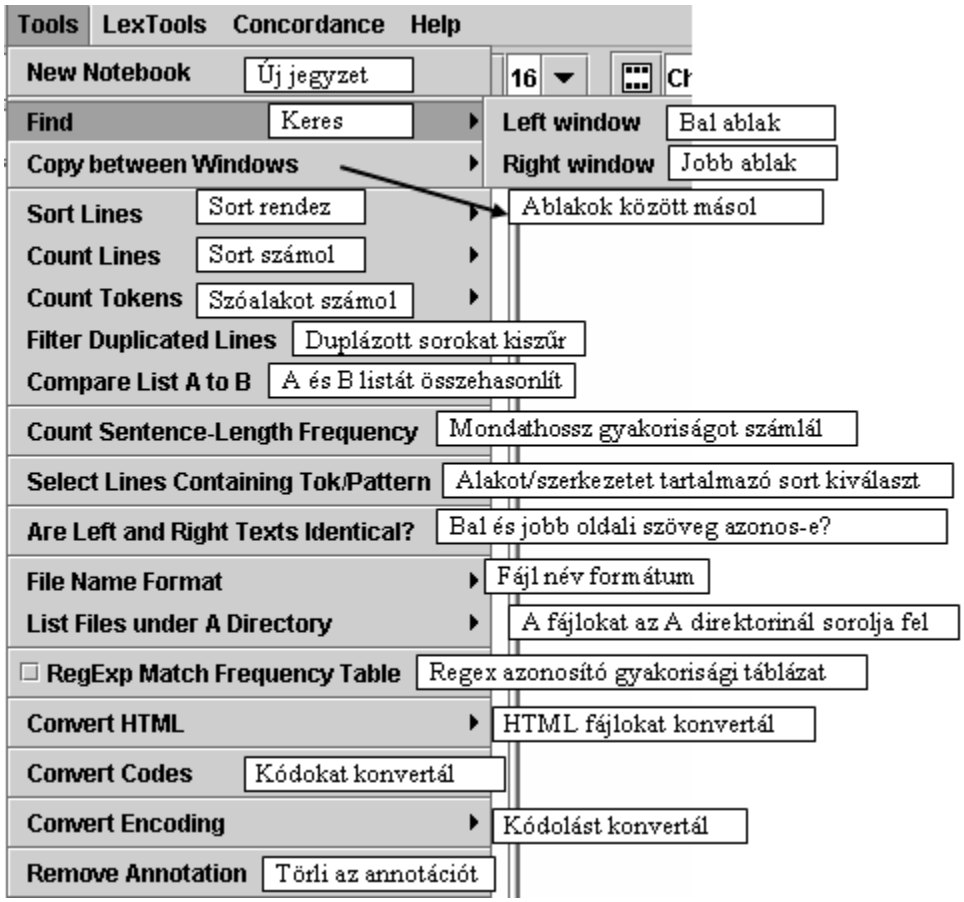
New left window	Új bal ablak
Open in left window	Bal ablakban kinyit
Save left window as	Bal ablak mentése másként
New right window	Új jobb ablak
Open in right window	Jobb ablakban kinyit
Save right window as	Jobb ablak mentése másként
Extract text from webpage	Honlapról szöveget kivon
Exit	Kilépés

37. ábra: Az *MLCT* File menüje

Első lépésként nyilván valamilyen szöveget kell megnyitnunk, tehát a Bal ablakban kinyit menüre kattintva a megfelelő fájlt kiválasztjuk. A fájl formátum választási lehetőségei a következők: egyszerű szövegfájl (txt), Latex dokumentum, HTML, XML és SGML dokumentum. A másik lehetőség a honlapról való szövegkinyerés. Ebben az esetben figyelni kell arra, hogy a számítógépre telepített tűzfal (firewall) a program internetre való kapcsolódását engedélyezze.

A View menüben a jobb és a bal ablakban a szövegnek az ablakhoz való méretezését állíthatjuk, valamint a háttér és előtér színeit. A két következő menü tartalmazza az igazi nyelvi elemzés utasításainak nagy részét, melyek közül több almenüt is tartalmaz. A Tools menü számos almenüje a művelet elvégzési helyének választási lehetőségét kínálja fel, így a jobb vagy bal oldali ablakot. Esetenként a művelet elvégzését, pl. a Duplázott sorokat kiszűr esetében, a fájlokban is felkínálja. A HTML fájlokat vagy egyszerű szövegfájllá vagy a mondat/bekezdés határokat bejelölt szöveggé alakítja. A Conevrt Encoding (átkódolás)[29] esetében csak bizonyos kombinációk választhatók a listából, pl. UTF 16-ról UTF 8-ra stb.

[29] A számítástechnikában a különböző nyelvekhez tartozó kódlapok feladata az, hogy az adott nyelv karaktereit megfelelően kódolják és jelenítsék meg. Ha például az internetes böngészés során olvashatatlanul jelenik meg egy szöveg, a kódolás állításával, azaz a helyes kódlap kiválasztásával korrigálhatjuk a hibát. Jelen esetben magát a kódlapot változtathatjuk meg.



38. ábra: A Tools menü pontjai

A fenti ábrán a menüpontok önmagukért beszélnek, így külön magyarázatot nem fűzünk ezekhez.

A LexTools menü neve is elárulja, hogy itt komolyabb lexikai jellegű eredményekre számíthatunk. Az első két menüpont a leghasznosabb az átlagos felhasználó számára. Az első, a Remove Punctuation Marks? (Eltávolítja az írásjeleket?), bekapcsolásával vagy kikapcsolásával meghagyhatjuk vagy eltávolíthatjuk az írásjeleket a szöveg vizsgálatakor. Az Extract N-grams (n-gramok kinyerése) pont almenüjéből 1-től 6-ig választhatunk. De mi is az n-gram (ejtsd: engrem)? Ha az 1n-gramet választjuk, akkor egy szólistát kapunk, ahol minden sorban egy szó szerepel. A 2n-gram esetében minden sorban két szó szerepel. Ezek a szavak a szövegben egymás mellett levő szavak. Nézzük meg ezt számokkal szemléltetve. Ha a szöveg szavait 1-től 10-ig terjedő számokkal helyettesítjük, akkor a 2n-gram a következőképpen néz ki:

12 23 34 45 56 67 78 89 910

39. ábra: 2n-gram számokkal szemléltetve

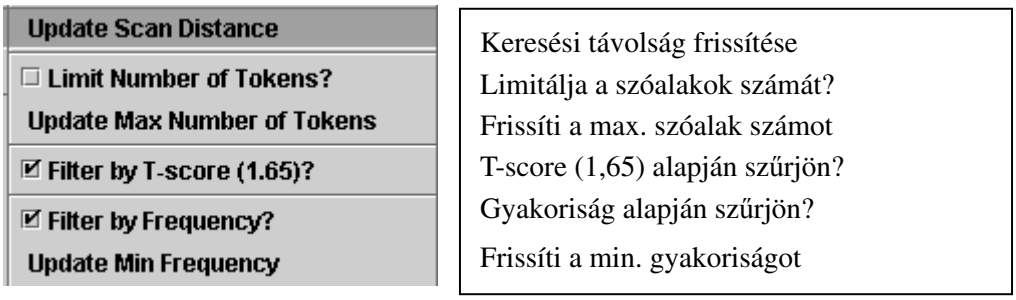
Ha 3n-gramet akarunk vizsgálni, akkor ugyanez a példa a következőre változik:

123 234 345 456 567 678 789 8910

40. ábra: 3n-gram számokkal szemléltetve

Mondathatárokat nem vesz figyelembe az ilyen elemzés, mely arra alkalmas, hogy olyan ismétlődő szerkezetekre hívja fel a figyelmet, amelyeket különben nem vennénk észre. Ennek kipróbálásához igen rövid fájl használatát javasoljuk, mert a program futtatása sok időbe telhet, és ezalatt azt is nehéz megállapítani, hogy a gép működik-e, egyáltalán érdemes-e várni. Az Extended Porter's Stemmer (Bővített Porter szótövező) csak az angol nyelv vizsgálatakor használható, hiszen az angol nyelv sajátosságainak megfelelően szótövekre „vágja" a szöveget. A következő két menüpont a kollokációk vizsgálatánál használható.

A Collocation Parameters (Kollokációk paraméterei) alpontjai a következők:



41. ábra: A Collocation Parameters almenüje

Ezek közül a T-score (ejtsd: tíszkór) magyarázatra szorul. Ez olyan statisztikai adat, amely megmutatja, hogy hogyan viszonyul egy-egy kollokáció tényleges előfordulása a valószínű előforduláshoz. Minél nagyobb ez a szám, annál biztosabb a kollokáció előfordulása. Az utolsó menüpont megértéséhez komoly ismeretek szükségesek, ezek meghaladják az átlag felhasználó szükségleteit.

42. ábra: Az Extract Collocates By statisztikai együtthatókat felkínáló alpontjai

A fenti listából csak egy elemet emelnénk ki, a Mutual Information-t, amelynél a vizsgált elem és kollokánsa arról adnak kölcsönösen információt, hogy tényleges együttes előfordulásuk hogyan viszonyul a várható előforduláshoz, feltételezve azt, hogy előfordulásuk esetleges. (Lásd McEnery & Wilson, 1996; Oakes, 1998; Ooi Vincent, 1998). Ha egy szó kollokációit különböző statisztikai számítások alapján készítjük, a listákon szereplő szavak vagy azok sorrendje más és más lesz. Ha a fenti statisztikai módszerek alaposabb megismerésére törekednek, Oakes könyvének 4. fejezetét (1998: 149–197) ajánlom további tanulmányozásra. Az angolul nem tudók a 171. oldalon találhatják meg az Ochiai-, McConnoughy-, Yule-, Fager/McGowan- és Kulczinsky-féle számítások matematikai képletét.

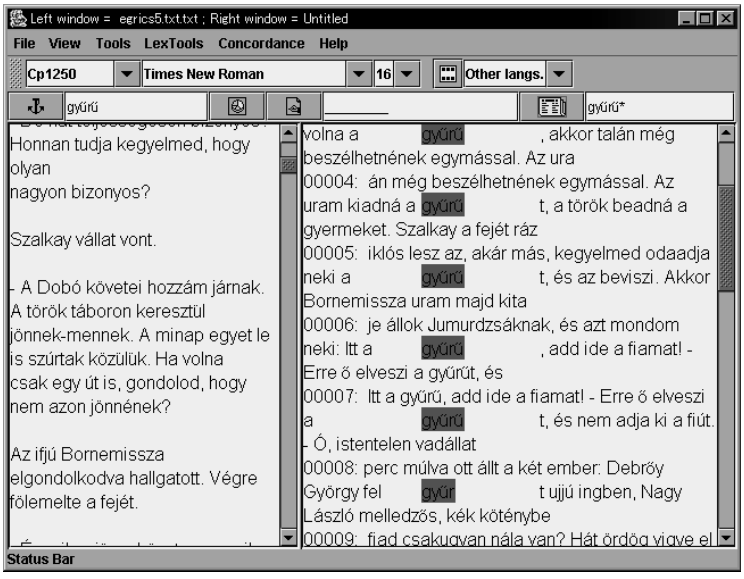
A Concordances (Konkordanciák) menüben frissíthetjük a paramétereket (szövegkörnyezet hosszát az első pontból), a másodikból a konkordanciák készítéséhez használni kívánt fájlokat választhatjuk ki, a harmadikkal a kiválasztást törölhetjük, és az utolsó pontnál a konkordanciák ábécérendbe való állításának módját választhatjuk ki: balra az első, második és harmadik szó, valamint jobbra az első, második vagy harmadik szó.

Még négy ikon és három szövegablak szorul magyarázatra, melyeket a következő ábrán láthatunk:

43. ábra: Kontroll ikonok és szövegablakok

Az első szövegablakba beírt kifejezést vagy szöveget a horgony ikonra való kattintással regexként keresi a bal ablakbeli szövegben. Az eredményt a jobb oldali ablakban láthatjuk. Amennyiben a Tools menü RegExp Match Frequency Table (Regex azonosító gyakorisági táblázat) almenüt bejelöljük, az eredményt rendezhető táblázat formájában is megtekinthetjük.

A kört ábrázoló ikon segítségével az első szövegablakba beírt kifejezést vagy szót a második ablakba beírt kifejezéssel „lecserélhetjük”, azaz helyettesíthetjük. Ha a második szövegablak üres, akkor az első szövegablakban szereplő kifejezést üressel helyettesíti vagy törli. Az eredményt a jobb oldali ablakban láthatjuk.



44. ábra: A programablak konkordanciákkal a jobb oldalon

A kéz ikon arra szolgál, hogy a második szövegablakban megadott módon megváltoztassa a keresett elemet, majd a megváltoztatott keresés eredményét kilistázza a jobb oldalon.

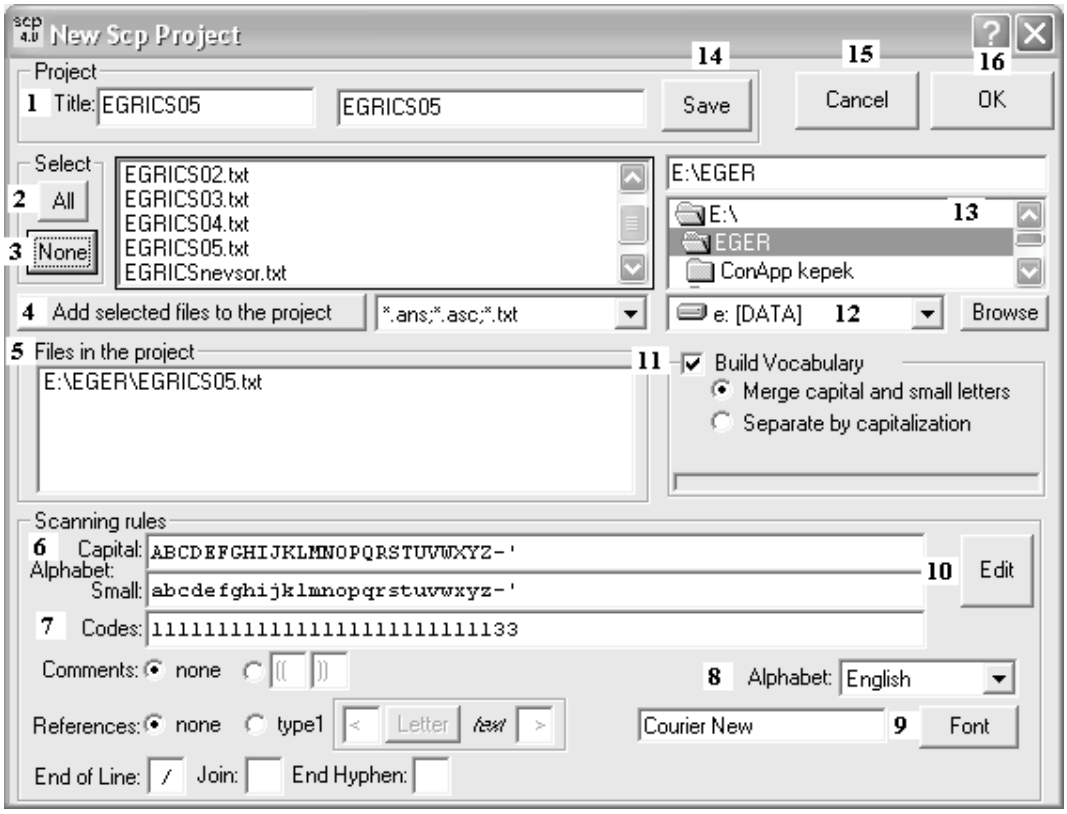
Az utolsó, könyv ikon a harmadik szövegablakba gépelt szöveg vagy regex konkordanciáit keresi ki a bal ablakban levő szövegből és a jobb oldali ablakban teszi láthatóvá. A harmadik szövegablakba gépelt kifejezés zöld színnel lesz feltüntetve a konkordanciákban. A 16. ábra jobb oldali ablakában láthatjuk a *gyűrű* szó konkordanciáját.

4.4.2. Simple Concordance Program SCP

Azért választottuk másodikként bemutatásra ezt a viszonylag nagyobb méretű programot, mert egy kis igazítással a magyar nyelvre „hangolható”, és majdnem a magyar ábécé szerinti listát leszünk képesek létrehozni segítségével. A problémát csak a két betűből álló kapcsolatok okozzák, így tehát a *cs*-vel kezdődő szavak a *cu*-val kezdődők előtt fognak szerepelni, és nem pedig utánuk. Az ékezetes betűk ebben a programban nem okoznak problémát, ha a projekt megkezdésekor a karakterkészletet kiegészítjük a magyar ékezetes betűkkel. Így tehát a magyar szavakat szóként fogja felismerni a program, míg más programok az ékezetes betűket a szavak keresésekor szóközként értelmezik.

A program indítása után megjelenő ablakban első lépésként a File menüből az Open, azaz Megnyit almenüt választjuk. Automatikusan a program mappája (scp32v407) nyílik ki és a program részeként letöltött, már kész scp kiterjesztésű projektek közüli választási lehetőséget kínálja fel. Ezek angol nyelvűek (2cities, gawain, lincoln), de kísérletezésre és tanulásra talán az angolul nem tudóknak is hasznosak lehetnek. Ha azonnal saját szöveggel akarunk dolgozni, akkor két lehetőségünk van. Vagy ugyanebből az ablakból folytatva a szokásos módon megkeressük a kívánt szövegfájlokat tartalmazó mappát, vagy pedig az Open menü helyett a New választásával azonnal egy új projektet kezdhetünk. Ha az előbbi megoldást választjuk, azaz a felkínált projektek helyett választunk saját fájlt, akkor ne felejtsük el a fájltípust átállítani szövegre, mert különben nem jelennek meg a listán a szövegfájlok. Ekkor ugyan még csak egy szövegfájlt választhatunk ki és nyithatunk meg, de a következő ablak kinyílása után a mappában levő összes szövegfájlt is kiválaszthatjuk a projekthez. Így javasoljuk, hogy vagy az összes szövegfájlt tegyék egy mappába, vagy legalábbis a projekthez használni kívántakat, még a program elindítása előtt. A kétféle kezdési mód ugyanahhoz az ablakhoz vezet, amit a 45. ábra mutat be. Az 1-es számmal jelölt szövegdobozba írhatjuk be a projekt nevét. Az ez alatt levő szövegdobozban látható az összes szövegfájl, ami a megnyitott mappában szerepel. Ha mindet ki akarjuk választani, akkor csak a 2-sel jelzett All gombra kell kattintani, és automatikusan az 5-ös számmal jelzett dobozba ugranak, amely a projektbe felvett fájlokat mutatja. Ha tévedtünk, akkor vagy az összes fájlt törölhetjük a projektből a 3-as gomb None megnyomásával, vagy az alsó dobozban levő nem kívánt fájlt, a nevére kétszer kattintva. Ha egyesével akarjuk a mappából kiválasztani a projektbe kerülő fájlokat, akkor ebben az esetben is duplán kattinthatunk a felső dobozban a nevére, vagy egy kattintással kijelöljük, és a 4-es gombot választjuk.

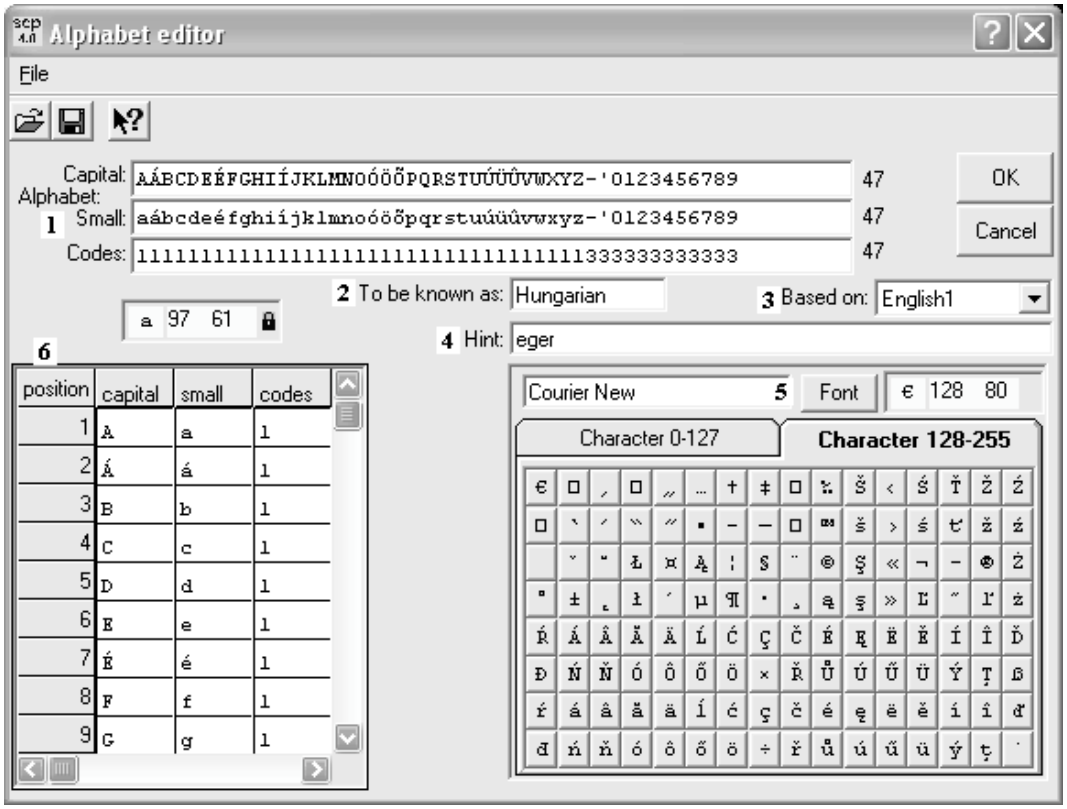
A szövegfájlok kiválasztása után a következő fontos lépés a projekthez szükséges karakterek és betűtípusok megadása. A 8-as számnál láthatjuk, hogy jelenleg angol nyelvre van állítva és a 6-os számnál két sorban szerepel az összes ebben a beállításban használt karakter, felül a nagybetűk, alul pedig a kisbetűk. A 7-es számnál kódokat látunk, melyek a számítógép számára adnak utasítást, hogy hogyan kezelje ezeket a karaktereket. Ha nem magyar, hanem más nyelvet használunk, például oroszt, németet, franciát, spanyolt, dánt, görögöt, izlandit, svédet, arabot, norvéget, hébert vagy katalánt, akkor csak a megfelelő nyelvet kell kiválasztanunk. Természetesen a számítógépünkön meg kell, hogy legyen a megfelelő nyelvi támogatás is. A Font (9-es gomb) megnyomásával egy újabb ablak nyílik meg, ahol kiválaszthatjuk a betűk típusát, stílusát és méretét a megfelelő írásrendszerrel együtt.



45. ábra: Új projekt létrehozása

Mivel a magyar nyelv nem szerepel a választható nyelvek között, elkerülhetetlen a betűkészlet saját kezű szerkesztése. A 10-es gombot megnyomva a 46. ábrán látható ablak nyílik meg. A szerkesztés legegyszerűbb módja, ha egy már létező, a magyar ábécéhez leginkább közel álló karakterkészletet egészítünk ki. Ebben az esetben az English 1-re esett a választás, mely a 3-as gombnál választható ki. Ezek után az 1-gyel jelzett szövegdobozban a megfelelő helyre kattintunk, ahol megjelenik a kurzor. A jobb alsó sarokban levő karaktertáblából az egérrel kiválasztjuk a megfelelő betűt, amely az egyessel jelzett szövegdobozban a kurzor helyén azonnal megjelenik. A nagy és kisbe-

tűket egyesével kiválasztva, ne felejtsük el a kódokat sem beírni! Minden beírt betű kódja 1 lesz. A sorok végén levő számoknak egyezniük kell. A 6-os számmal jelzett oszlopok a fentebb levő szövegdobozzal egyező információt mutatnak, de talán könnyebb itt észrevenni a hibát, mint a felsorolásban. A 2-vel jelzett szövegdobozba írjuk az általunk választott nevet a karakterkészlet számára. A 3-at üresen is hagyhatjuk, vagy valami utalást írhatunk bele. Végül ne felejtsük el az OK gombot megnyomni. Ha valamit elvétettünk, a program egy ablak megjelenésével ezt jelzi. Ha nincs probléma, akkor a 45. ábrán látható ablakhoz jutunk vissza.

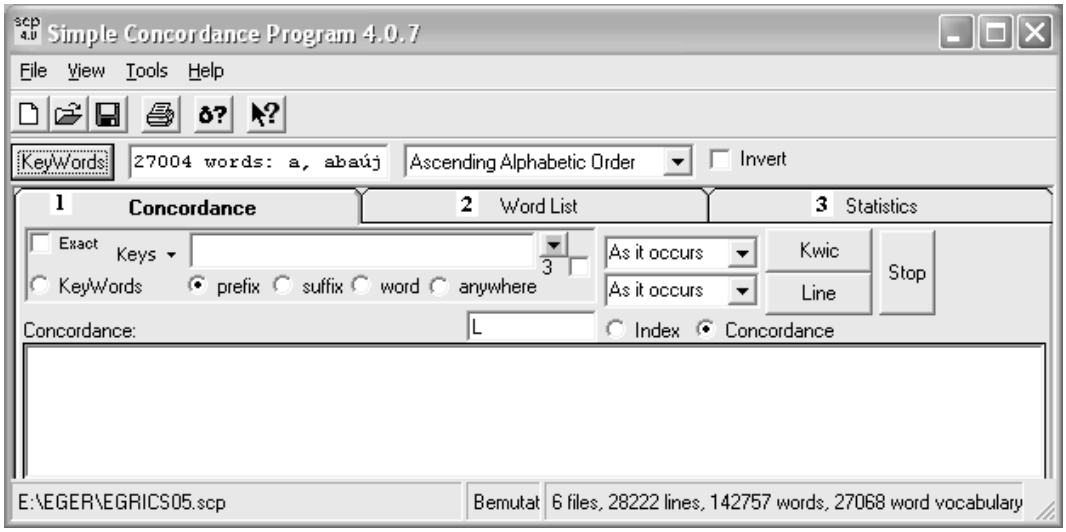


46. ábra: A karakterkészlet szerkesztőablaka

Ha a szövegfájlban nincs magán a szövegen kívül semmilyen megjegyzés vagy referencia, akkor a kódok alatt szereplő választási lehetőségeket az eredeti állapotban kell hagyni.

A 11-es gombnál a Build Vocabulary (Szószedet létrehozása) bejelölésével a program a projekt szószedetét automatikusan elkészíti és tárolja. A közvetlenül alatta levő gomb választásával a nagybetűs és kisbetűs szavakat nem sorolja fel külön, hanem azonosnak tekinti, tehát a *kovács* mesterség és a *Kovács* tulajdonnév között nem tesz különbséget. Ha az alatta levőt választjuk, akkor szétválasztja a kis- és nagybetűs szavakat. Ha mindent megfelelően kiválasztottunk, akkor a 12-es számnál kiválasztjuk, hogy melyik merevlemezen, a 13-asnál pedig, hogy melyik mappába akarjuk a projektet elmenteni. Az OK gombra kattintva megjelenik egy ablak, amely azt a kérdést teszi fel, hogy menteni kívánjuk-e a projektet. Mindenképpen ajánlatos elmenteni, legalábbis addig,

amíg jobban meg nem ismerjük a program működését. Az igenre kattintva újabb ablak nyílik, ahol menthetjük a projektet.



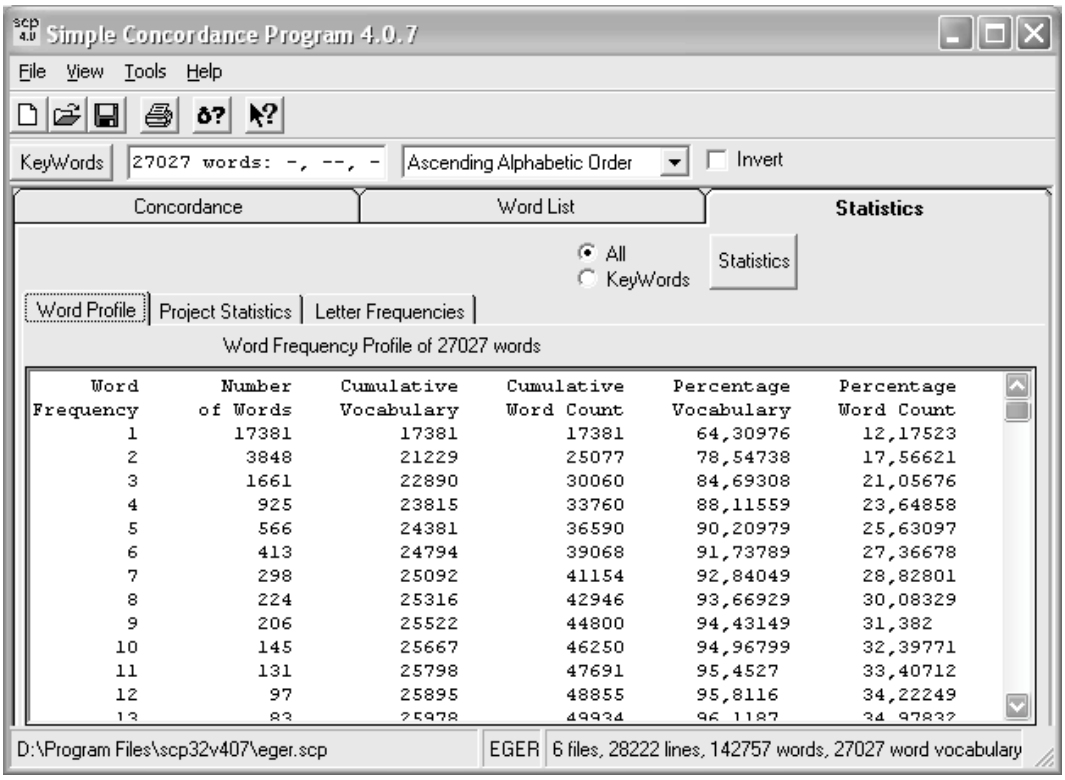
47. ábra: A projekt kezdő ablaka

A projekt mentése utáni kezdőablakot a 47. ábra szemlélteti. Három nagy egységet láthatunk: 1. Konkordancia; 2. Szólista; és 3. Statisztika felirattal. A fenti ábra a Konkordancia választási lehetőségeit nyújtja. Mielőtt azonban rátérnénk ennek tárgyalására, figyeljük meg, hogy az ablak alján már most látható néhány fontos statisztikai adat. A projekthez vezető elérési út mellett a projektben szereplő fájlok, sorok, szövegszó és szóalakok száma látható. A Keys melletti szövegdobozba írjuk be a keresett szót, és az alatta levő sorban válasszuk a word kifejezést, majd kattintsunk a Kwick gombra. Ekkor a projekt szövegéből a beírt kulcsszó Kwick formátumban levő előfordulásai az alsó nagy szövegdobozban jelennek meg. Így ha ide a *török* szót írjuk be, akkor csak azokat az előfordulásokat választja ki a program, amelyben a *török* alak áll, de a toldalékos szavakat, mint a *töröknek,* már nem. Ha ezekre is kíváncsiak vagyunk, akkor a *török*öt, mint prefixumot kell keresni, annak ellenére, hogy a magyar nyelvben ez nem prefixum. (Erre azért van szükség, mert a program nem a magyar nyelv sajátosságait figyelembe véve készült, hanem az angol nyelv logikáját követi.) Még egy választási lehetőség van. Az anywhere, azaz bárhol választásával is erre az eredményre juthatunk e szó esetében. Azonban ez nem mindig célravezető, például ha az *ér* szót bármilyen előfordulásban keressük, nagyon valószínű, hogy sok olyan szó jelenik meg, amelyben az *ér* olyan módon szerepel, mint például a *denevér* szóban. A konkordanciák megjelenéséig akár egy teljes perc is eltelhet, a fájl méretétől és a számítógép kapacitásától függően, ezért először érdemesebb kisebb fájlon kísérletezni. Ha a Kwick helyett a Line gombot választjuk, akkor a keresett szó nem a sor közepén fog megjelenni, hanem a sor bármely részén, hiszen itt azt a teljes sort láthatjuk, amelyben a keresett szó szerepelt.

A 2-vel jelzett Word List esetében nem sok teendőnk van. A szólista négyféleképpen jeleníthető meg: balra vagy jobbra igazodó oszlopokban, sűrítve, és egy oszlopban. A Layout címszó alatt ilyen sorrendben találjuk meg ezeket. A könnyű áttekinthetőség

érdekében javasoljuk az egy oszlopot. Ezek után már csak a Word List feliratú gombot kell megnyomni, és a szavak ábécé sorrendben, előfordulási számukkal együtt megjelennek. Ha más sorrendben szeretnénk látni az adatokat, például először a leggyakrabban előforduló szót, és csökkenő sorrendben a többit, akkor ezt a Word List felirat feletti legördülő ablakból választhatjuk ki.

A Statistics címszó alatt elég, ha csak egyet kattintunk a Statistics gombra, és automatikusan a 48. ábrán mutatotthoz hasonló információt kapunk.



Word Frequency	Number of Words	Cumulative Vocabulary	Cumulative Word Count	Percentage Vocabulary	Percentage Word Count
1	17381	17381	17381	64,30976	12,17523
2	3848	21229	25077	78,54738	17,56621
3	1661	22890	30060	84,69308	21,05676
4	925	23815	33760	88,11559	23,64858
5	566	24381	36590	90,20979	25,63097
6	413	24794	39068	91,73789	27,36678
7	298	25092	41154	92,84049	28,82801
8	224	25316	42946	93,66929	30,08329
9	206	25522	44800	94,43149	31,382
10	145	25667	46250	94,96799	32,39771
11	131	25798	47691	95,4527	33,40712
12	97	25895	48855	95,8116	34,22249
13	83	25978	49934	96,1187	34,97832

48. ábra: Statisztikai adatok

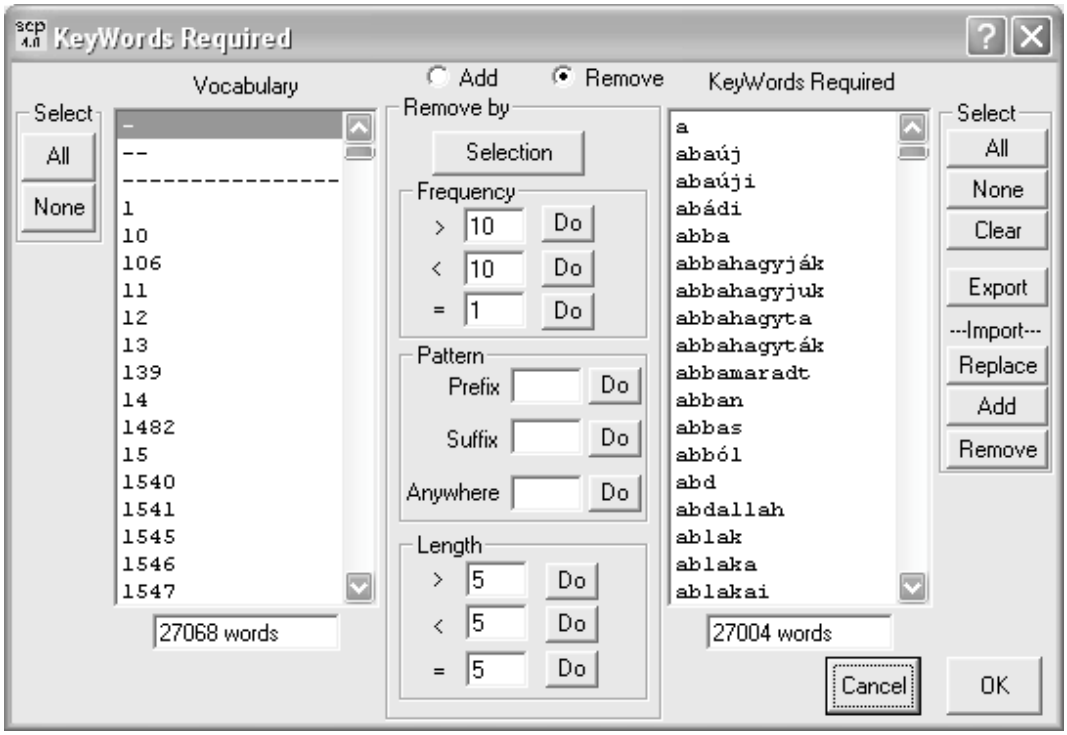
Az első oszlop a szavak gyakoriságát mutatja, tehát az első sorban az egyszer előforduló szavak szerepelnek. A második oszlop az egyszer előforduló szavak számát mutatja. Ebben a projektben 17 381 szó szerepel csak egyszer a szövegben. A harmadik oszlopban azt láthatjuk, hogy ez az eddigiekkel együtt összesen hány szót tesz ki. A harmadik oszlop második sora a második oszlop első és második sorának összege, tehát az egyszer és kétszer előforduló szavak szóalakjának számát mutatja. A negyedik oszlop második sora azt mutatja, hogy az egyszer és kétszer előforduló szavak hány szövegszót jelentenek. Az ötödik oszlop a szóalakokhoz viszonyított arányt, a hatodik pedig a teljes szövegszóhoz viszonyított arányt mutatja. Ha közelebbről megnézzük, tanulságos, hogy a 27 ezres szókészletből több mint 17 ezer szó, kb. 64%, csak egyszer fordul elő. Természetesen itt a toldalékos alakokat külön számítottuk.

A projektstatisztikában nem kapunk túl sok plusz információt, és ez is inkább technikai jellegű.

```
Analysis based on the whole vocabulary   Az elemzés a teljes szószedetre épül
Total vocabulary = 27068 types           Teljes szókészlet = 27068 szóalak (típus)
Project wordcount = 142757 tokens        Projekt szószáma = 142757 szó
Types/tokens = 0,18960892                Szóalak/összes szó = 0,18960892
Types/sqrt(tokens) = 71,64031082         Szóalak/
Yule's k = 154,35324502                  Yule
```

49. ábra: A projektstatisztika adatai

Az utolsó rész betűgyakoriságot mutat. Ebből tudhatjuk meg, hogy az egyes betűk hányszor szerepelnek a teljes szövegben. Talán még egy fontos dolgot kell megemlítenünk. A 47. és 48. ábrán levő ikonsor alatt láthatunk egy KeyWords feliratú gombot. Ha erre kattintunk, a következő ablakot láthatjuk:



50. ábra: Kulcsszavak szerkesztése

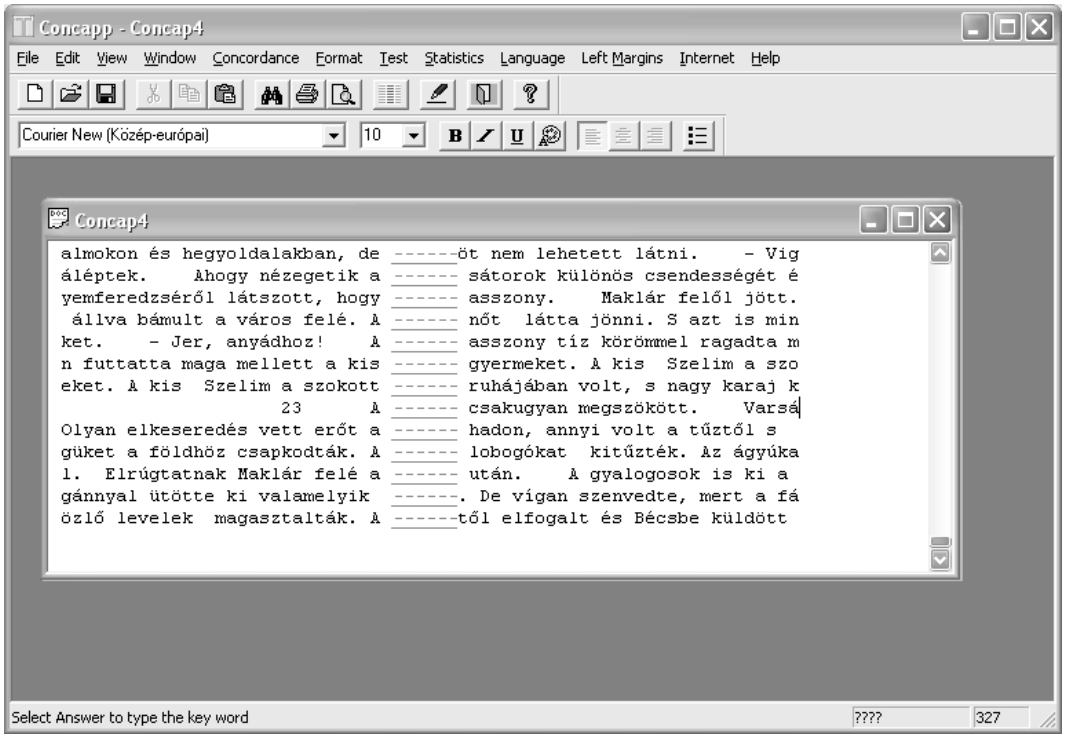
Az ablak két jól elkülöníthető, bal és jobb oldali listát tartalmaz. A bal oldalon látható az összes egység, amelyet a számítógép szóként értelmez. A szó meghatározása a számítógép számára az, amit két oldalról szóköz határol. Jól látható a bal oldali oszlopnál, hogy a gondolatjelet és a számokat is szónak tekinti a program. Ha azonban a felhasználó ezeket nem tekinti annak, akkor nincs rájuk szükség a szótárban. Az ilyen felesleges „szavak” szűrésére szolgál ez az ablak. Hozzá lehet adni, vagy törölni lehet szavakat. Ebben az esetben látható, hogy a jobb oldali oszlopban már nem szerepelnek a számok, töröltük őket. Automatikusan is törölhetők elemek, például a minimális gyakoriság megadásával. Az ablak jobb oldalán látható, hogy a kulcsszavakat exportálni és importálni is lehet.

A program menüi meglehetősen egyszerűek. Talán a Tools menü Apply a Stop List almenüjéről érdemes még szólnunk. Mi is a *stop list*? Minden nyelvben vannak olyan szavak, amelyek nagyon gyakran előfordulnak, de igazából nem vagyunk rájuk kíváncsiak vizsgálatuk során, mert nem jelentéshordozók. Az angolban ilyen például a névelő, a prepozíciók stb. Annak érdekében, hogy ezek ne „szennyezzék" a listánkat, egyszerűen kikapcsolhatók, tehát nem jelennek meg a gyakoriság vizsgálatakor, ha ezt úgy kívánjuk.

4.4.3. ConcApp

E program indítása után arra vigyázzunk, hogy ne MS-DOS-os szövegfájlként, hanem DOS dokumentumként nyissuk ki a vizsgálni kívánt fájlt. Ez a program angol, kínai és japán szövegek vizsgálata céljából készült, így bizonyos funkciók nem működnek ideálisan a magyar nyelv esetében. Ezt a programot nem írjuk le olyan részletességgel, mint az előző kettőt, hanem olyan funkcióját említjük elsősorban, amely a többiben nem található meg.

A program indítása után megjelenik egy információs ablak, amely kattintásra eltűnik. A vizsgálandó fájl kiválasztása után a menüsorból válasszuk ki a Test menüpontot. A New választásával új ablak jelenik meg, ahol begépelhetjük a kívánt kifejezést vagy szót. Az OK gombra való kattintással a következő ablakot kapjuk:



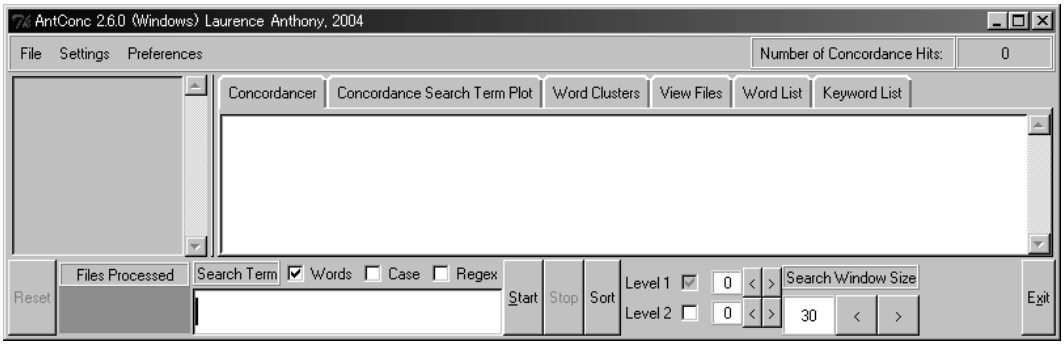
51. ábra: Teszt funkció a *ConcApp* programban

A keresett szót vonal helyettesíti, ami arra ad alkalmat, hogy a diák a keresett szót kitalálja. E funkció segítségével könnyen lehet játékos teszteket készíteni. Ezt a funkciót találtuk a leghasznosabbnak ez esetben.

4.4.4. AntConc

Ez a program sajnos nem képes a magyar ékezetes betűket értelmezni, ha szavakat keresünk. Így tehát szólistát nem tudunk ezzel készíteni. Ennek ellenére vannak olyan funkciók, amelyeket most is jól lehet használni. Mivel azonban a programok állandó fejlesztés alatt állnak, elképzelhető hogy mire e könyv az olvasó kezébe kerül, ez a probléma megoldódik, és így az itt leírtak nem fedik majd teljesen a valóságot. (Ez lenne a jobbik eset.)

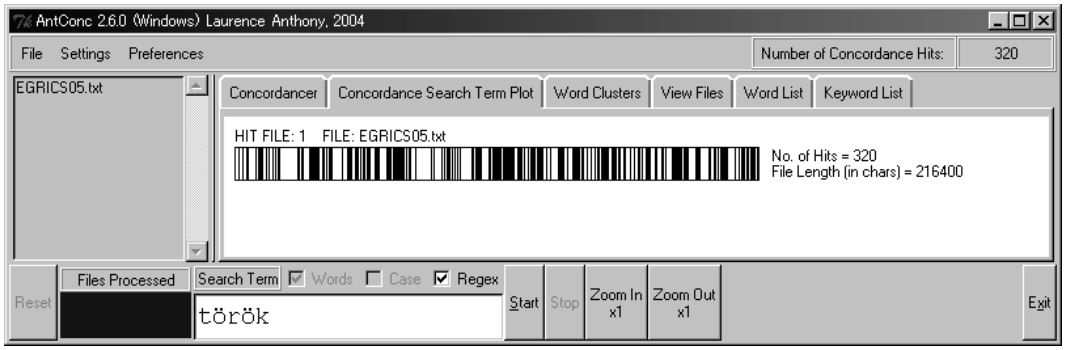
A program kezelőfelülete jól áttekinthető (52. ábra), a keresett szót vagy kifejezést az alsó szövegablakba kell beírni. A kurzor a program indításakor automatikusan ott villog, így nehéz eltéveszteni. Közvetlenül e fölött található a keresés módja, ahol a szóra való keresést vagy a regexet jelölhetnénk be, ha a program kezelni tudná az ékezetes betűket. Így csak a regex vezet eredményre. Természetesen a keresés megkezdéséhez meg kell nyitnunk egy fájlt, amit a szokásos módon a fájl menüből tehetünk meg.



52. ábra: Az *AntConc* kezelőfelülete

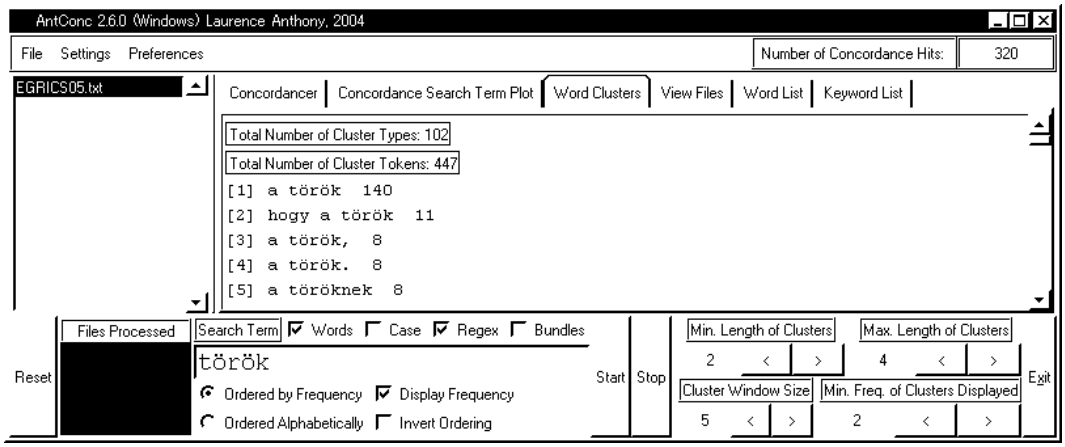
A keresett kifejezés konkordanciái a középső, nagy szövegablakban jelennek meg, a keresett kifejezés kék színnel van jelölve. A konkordanciákat a Level 1 és Level 2 gomb melletti szám, a képen 0, megváltoztatásának segítségével lehet úgy rendezni, hogy a különböző szerkezetek könnyen észrevehetőek legyenek. Ha például az 1R-t választjuk és a Sort gombra kattintunk, akkor a keresett szó melletti jobbra eső szavakat fogja ábécé szerinti sorba rendezni. Nincs megszabva, hogy hányas számra állíthatjuk ezt a funkciót, de két vagy háromnál többre nemigen érdemes állítani. A képen 30-ra állított szám változtatásával az ablakban egy sorban szereplő szövegmennyiséget lehet változtatni.

A keresés eredményét megtartva, a Concordance melletti fülre kattintva a vásárlásból jól ismert vonalkódra hasonlító képet láthatunk (53. ábra), mely azt szemlélteti, hogy a keresett kifejezés a szövegben hol helyezkedik el. Minden egyes előfordulást egy függőleges vonal jelez, így a sűrű vonalak azt mutatják, hogy ezek egymáshoz igen közel vannak.



53. ábra: A keresett szó elhelyezkedése a szövegben

A Word Clusters fülre való kattintással újabb formában vizsgálhatjuk a kívánt szót vagy kifejezést. Ne feledjük azonban, hogy most is csak regexként kereshetünk. Mire jó ez a funkció? Az MLCT program leírásakor említettük az n-gram keresési módját, és hogy ez meglehetősen igénybe veszi a számítógép kapacitását. Az AntConc e funkciója arra használható, hogy a keresett kifejezést az általunk megadott minimum és maximum szócsoportokban kigyűjtse előfordulásuk számával együtt.



54. ábra: Szócsoport-keresés

A keresett szó alatti részen lehet beállítani, hogy kívánjuk-e a gyakoriságot a képernyőn látni vagy sem, hogy milyen sorrendben, növekvő vagy csökkenő gyakoriság szerint, ábécé szerinti vagy fordított sorrendben akarjuk-e látni a szócsoportokat.

A View Files az eredeti szövegfájlt mutatja, a keresett szóra lehet a szövegben ugrani az előző (Previous Hit) és a következő (Next Hit) gombok segítségével. Itt meg szeretnénk jegyezni, hogy a konkordanciák nézetből is a szövegre lehet ugrani, ha a keresett szó fölé vitt egérmutató átváltozik kézzé, és ekkor a szóra kattintunk.

A Word List, azaz szólista funkció használhatatlan a magyar nyelv esetében, ha ékezetes betűk is szerepelnek a szövegben.

A Keyword List használatához egy referenciakorpuszra is szükség van. Mivel a program a magyar ékezetes szavakat nem értelmezte szavakként, ezt a funkciót ki sem próbáltuk.

4.5. Összefoglalás

Manapság nem az információhiány a probléma, hanem inkább az, hogy a rengeteg rendelkezésre álló, bennünket időnként elöntő áradatot nehéz befogadnunk. A számítógépek fejlődésével és az internet elterjedésével ez nem csak a magánéletre igaz. Az internet és az elektronikus könyvek korában a nyelvi és nyelvészeti vizsgálatokhoz rendelkezésre álló adathalmaz hatalmas. Minden országban igyekeznek minél nagyobb nemzeti korpuszokat létrehozni a korpusz alapú nyelvészeti leírások érdekében. Ebben a fejezetben először a korpuszkészítéskor használt programokat említettük meg, de nem írtuk le ezeket részletesen, hiszen elsősorban nem az annotált korpusz készítésére akarjuk biztatni az olvasót, hanem már „kész" korpuszok használatára, vagy egy saját szövegfájlokból álló korpusz használatára.

A korpuszok elemzésének egyik alapvető módja a konkordanciák elemzése. A konkordanciaprogramok általános működésének bemutatása után röviden áttekintettük a korai konkordanciaprogramokat, amelyek még ma is elérhetők az interneten.

Az internetes felületen futó konkordanciaprogramok a kimondottan magyar nyelvre készültek kivételével nem használhatók jól a magyar nyelv esetében az ékezetes betűk miatt. Így a fejezet nagy részét annak szenteltük, hogy négy, az internetről ingyenesen letölthető programot több-kevesebb részletességgel bemutassunk (MLCT, SCP, ConcApp és AntConc). Azért éreztük szükségét annak, hogy több programot is bemutassunk, mert egyikük sem a magyar nyelv speciális igényeinek figyelembevételével készült. Így a különböző programok különböző funkciói egymástól eltérő módon működnek, alkalmasabbak vagy kevésbé alkalmasak a magyar szövegek vizsgálatára.

Javasoltuk, hogy a programleírások olvasásával egy időben az olvasó is próbálja ki az adott program működését a saját számítógépén, saját szövegfájlja segítségével. A következő fejezetben utalunk a programok különböző funkcióira, de végrehajtásuk módját nem ismertetjük újból.

5. KORPUSZNYELVÉSZETI MÓDSZEREK AZ OKTATÁSBAN

5.1. Bevezetés

Jól ismert tény, hogy a tudományos eredmények megjelenése és az oktatásba, valamint tankönyvekbe való bekerülése között általában hosszú évek telnek el. Jóllehet az első elektronikus korpusz megjelenése óta 40 év telt el, a korpusznyelvészet igazából csak körülbelül 10 évvel ezelőtt kezdte igazán éreztetni hatását az oktatásban. Az utóbbi néhány évben azonban a *korpusz* és *korpusznyelvészet* kifejezések a nemzetközi nyelvtanári körökben és konferenciákon kimondottan gyakran hallott kifejezésekké váltak. Ezek alapján úgy véljük, hogy egy jól felkészült nyelvtanárnak, függetlenül attól, hogy anyanyelvi vagy idegen nyelvi oktatással foglalkozik-e, ismernie kell majd a korpusznyelvészet alapjait és az erre építő pedagógiai módszereket, és tanítási anyagok készítését.

Több akadálya is van ezen a téren a nyelvtanárok önképzésének, nem csak Magyarországon, de más országokban is. Az első akadály az angol nyelv lehet, hiszen a szakirodalom nagy része angol nyelven jelenik meg, még abban az esetben is, ha más nyelvről készült a tanulmány. Az angolul jól tudók számára az internet sok információval szolgálhat, de igazából nem pótolhatja egy korpusznyelvészetről szóló könyv szerepét, és bizony ezek ára igen borsos is lehet. Sokak számára a számítógép az akadály, hiszen már a puszta említése is ellenszenvet vált ki. Bízunk benne, hogy e fejezet elolvasása után kíváncsiságuk legyőzi majd esetleges negatív érzéseiket, és kipróbálnak legalább néhányat az ajánlott gyakorlatok közül.

A fejezet célja az, hogy megfelelően tájékoztassa az olvasót arról, hogy milyen más forrásokból szerezhetnek be információt (konferenciákról, publikációkból) a korpusznyelvészet oktatásban való felhasználásáról, és hogy milyen lépéseket javasolunk megtenni annak érdekében, hogy a saját tanítási gyakorlatukban is használhassák ezeket a módszereket. A különböző feladatok és gyakorlatok bemutatásakor nem a teljességre törekedtünk, hanem gondolatébresztőnek szántuk azokat.

5.2. Konferenciák és publikációk

Az egy-egy kutatási területre összpontosító konferenciák és publikációk megjelenése bizonyítja azt, hogy az adott szakmában az e téma kutatásával foglalkozók száma és a kutatás fontossága indokolja azt, hogy önálló konferenciát rendezzenek. Így a korpusznyelvészet oktatásban való felhasználását vizsgáló kutatások iránti érdeklődés megnövekedésének eredményeként, számos e témának szentelt, vagy e témának is megfelelő fórumot biztosító konferenciákat rendeztek meg. A Teaching and Language Corpora

('Oktatás és nyelvi korpuszok') konferenciát 1994 óta kétévente rendezik meg. Ezt követte két másik konferencia a Practical Applications of Language Corpora ('Nyelvi korpuszok gyakorlati alkalmazásai') és a Corpus Use and Learning to Translate ('Korpuszhasználat és a fordítás tanulása') 1997-ben. Az előbbit szintén kétévente rendezik, ami azt jelenti, hogy a konferenciák váltakoznak ugyan, de minden évben van legalább egy olyan konferencia, ahol a korpuszok oktatásban való használatával kapcsolatos új kutatások eredményeit közzé lehet tenni. A fordítás tanítására vonatkozó konferencia hosszabb időközönként kerül megrendezésre, ez valószínű, hogy speciális jellegéből fakad. Az 24. táblázat összefoglaló képet ad a korpuszok oktatásban való felhasználásával foglalkozó konferenciákról. A táblázatot Federico Zanettin honlapjáról (http://www.federicozanettin.net/sslmit/cl.htm) vettük, és ezt egészítettük ki.

A konferenciákon elhangzott előadásokból a legtöbb esetben könyvet is készítettek, amelyek megvásárolhatók. E könyvben már sokszor elhangzott frázist kell itt megismételnem: a cikkek ezekben is angol nyelven jelentek meg. Számos előadás anyaga vagy annak részei az előadók vagy a konferencia honlapjáról is letölthetők. Így például a http://www-gewi.kfunigraz.ac.at/talc2000/Htm/index1.htm címről, a TaLC 2000 honlapjáról az előadások absztraktjai és egyes poszter előadások, valamint számos PowerPoint előadás még most is letölthető. Ha az előadást nem is pótolhatják ezek, számos hasznos információ és számadat segíthet a további anyagok felkutatásában.

TaLC (Teaching and Language Corpora)	**PALC (Practical Applications of Language Corpora)**	**CULT (Corpus Use and Learning to Translate)**
• TaLC 1994 (**Lancaster, UK**) • TaLC 1996 (**Lancaster, UK**) • TaLC 1998 (**Oxford, UK**) • TaLC 2000 (**Graz, Austria**) • TaLC 2002 (**Bertinoro, Italy**) • TaLC 2004 (**Granada, Spain**)	• PALC 1997 (**Lodz, Poland**) • PALC 1999 (**Lodz, Poland**) • PALC 2001 (**Lodz, Poland**) • PALC 2003 (**Lodz, Poland**) • PALC 2005 (**Lodz, Poland**)	• CULT 1997 (**Bertinoro, Italy**) • CULT 2000 (**Bertinoro, Italy**) • CULT 2004 (**Barcelona, Spain**)

25. táblázat: A korpuszok oktatásban való használatával foglalkozó konferenciák

A PALC 2005 rövidítésében ugyan nem változott, de nevében kisebb módosítás történt, így idén először a Practical Applications in Language and Computers ('Gyakorlati alkalmazások a nyelvben és a számítógépek') címmel tartják**,** melynek általános témája az informatika fejlődésének eredményei és azok felhasználásának kapcsolata a nyelvi és nyelvészeti területeken. Az előadások elfogadásának előfeltétele, hogy korpuszvizsgálatra épüljenek, és a 12 ajánlott témából három kimondottan az oktatásra vonatkozik – ezek a következők:

- idegen vagy második nyelv elsajátítása,
- nyelvtanítási anyagok és nyelvi korpuszok,
- nyelvtanítás és tanulói korpuszok.

Természetesen a konferenciakiadványok mellett számos könyvet és cikkgyűjteményt szenteltek a korpuszok oktatásban játszott szerepének és az eredmények bemutatásának. Itt szeretnénk megemlíteni néhány olyan kötetet, amely átfogó képet ad. Elsőként jelent meg a *Teaching and Language Corpora* ('Oktatás és nyelvi korpuszok') (1997), melyet a *Learner English on Computer* ('Tanulói angol nyelv számítógépen') (S. Granger, 1998) és a *Rethinking Language Pedagogy from a Corpus Perspective* ('A nyelvpedagógia átgondolása korpuszszempontok szerint') (Lou Burnard & McEnery, 2000) követtek. A *Small Corpus Studies and ELT: Theory and Practice* ('Tanulmányok a kisméretű korpuszokról és az ELT: Elmélet és gyakorlat') (Ghadessy *et al.*, 2001), *Learning with Corpora* ('Tanulás korpusszal') (Aston, 2001), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* ('Számítógépes tanulói korpusz, második nyelv elsajátítása és idegennyelv oktatás') (Sylviane Granger *et al.*, 2002), és a *How to Use Corpora in Language Teaching* ('Hogyan használjunk korpuszokat a nyelvtanításban') (J. M. Sinclair, 2004) a legfrissebb kötetek ebben a sorban. Természetesen számos más könyvben is helyet kaptak az ilyen jellegű tanulmányok, de ezek a kötetek kizárólag e témával foglalkozó cikkeket tartalmaznak. A címek tulajdonképpen már el is árulják, hogy mennyire sokféleképpen lehet a korpuszokat felhasználni. A kötetekben szereplő cikkek címét látva nyilvánvaló, hogy a nyelvtanítás és a nyelvészet minden területén alkalmazhatók a korpuszok. Természetesen az egyetemi szintű szakképzés esetében, tehát a nyelvészek, nyelvtanárok, fordítók, tolmácsok esetében magát a korpusznyelvészetet is tanítják, amely nélkül nehéz lenne elképzelni a modern szakemberképzést. De még az egyetemi oktatás keretében is a szaktantárgy szerep mellett a nyelvtanulást segítő funkcióját is megtartja.

Mint láthatjuk, több ezer oldal jelent már meg a korpusznyelvészet szerepéről az oktatásban, így a fejezet következő részében néhány példával illusztrálva igazából csak ízelítőt adhatunk, hogy milyen tanítási anyagokat készíthetünk segítségével.

5.3. Számítógéppel és nélküle

A korpusznyelvészet oktatásban való használatának ideális módja az lenne, ha a diákok maguk kutathatnának a nyelvi szintjüknek és tanulási céljaiknak megfelelő korpuszokban. Ehhez persze szükség lenne megfelelő korpuszokra, programokra, számítógépes teremre, az ezekhez jól értő nyelvtanárra és a számítógépet gond nélkül kezelni képes, motivált és érdeklődő diákokra. A rideg valóság valószínűleg az, hogy magunknak kell megfelelő korpuszt összeállítani a diákok számára, ha azt kívánjuk, hogy ők maguk is használhassák ezeket. A megfelelő programok drágák, így jó, ha a tanár gépére meg tudják venni. A számítógépes termet a legritkább esetben használhatja halandó nyelvtanár nyelvórák tartására, hiszen az „nem arra való". Talán a diákok sem biztos, hogy olyan lelkesek és kíváncsiak, hogy ilyen új „játékokat" kipróbáljanak, hiszen igazából csak a vizsgán szeretnének már túl lenni. Talán eltúloztam? Valószínű. De biztos, hogy minden tanár helyzete és diákjai különbözőek, így általános érvényű tanácsot nem adhatunk, mindenkinek önmagának kell megtalálnia a legmegfelelőbb megoldást az adott helyzetben.

Ha minden tárgyi feltétel adott, akkor is szükségszerűnek látszik egy bizonyos fokozatossági elvet követni. Ez vonatkozik a tanárokra és a diákokra is. A tapasztalat azt mutatja, hogy a legtöbb tanár először a korpusz alapú könyvek és CD-ROM-ok használatával kezd megismerkedni, majd ezután magukkal a meglévő korpuszokkal, mégpedig olyan célból, hogy saját kérdéseire vagy diákjaiéra választ kapjon. Ilyen lehet például az, hogy „Mi a különbség X és Y között?", vagy „Mikor használjuk A-t és mikor B-t, ha majdnem ugyanaz a jelentésük?". A „késztermékek" alaposabb megismerése után érdemes csak saját anyagok készítésével próbálkozni. A korpuszok esetében ez az interneten elérhető korpuszokat jelenti, mert ezekhez viszonylag könnyen kezelhető keresőprogramok is tartoznak. A magyar nyelv esetében a Magyar Nemzeti Szövegtár (MNSZ) használata lehet az első lépés. Az angol nyelv esetében már sok más „késztermék" is létezik, amelyek korpuszelemzésekre épültek. Ezek rendszeres használatával jobban megismerhetjük, hogy milyen jellegű feladatok léteznek, és azok mintájára készíthetünk majd később saját tanítási anyagokat. A legegyszerűbb, ha a feladatok típusát követjük, és csak a tartalmukat változtatjuk meg először.

A harmadik lépés tehát a saját anyag készítése és azok használata a tanórákon. Ehhez az előző fejezetben leírt programok használatát javasoljuk. A saját kutatások eredményeit is fel lehet használni feladatlapok készítéséhez, de a kiválasztott konkordanciákhoz is készíthetünk feladatsort.

Azt tartjuk ideálisnak, ha a diákok maguk használják a korpuszt, de ehhez megfelelő korpuszra van szükség, hiszen míg az előzőekben minden a tanár szűrőjén keresztül került a diákok kezébe, ebben az esetben a diák és a korpusz között közvetlen kapcsolat áll fenn. A diákok használatára készült korpuszok kisebb méret esetén is megfelelőnek, sőt talán kívánatosnak is tűnnek. Az információáradat vagy érthetetlen adathalmaz nemhogy segítené a tanulást, hanem hátráltatja azt.

5.4. A késztermékek

Az javasoltuk, hogy első lépésként ezekkel ismerkedjenek meg, ha lehet. Az angol nyelv bővelkedik ezekben, de a többi nyelv igen szegényes e téren. Így elképzelhető, hogy sokaknak át kell ugraniuk ezt a lépést, és bele kell vetniük magukat a korpuszok használatába. Bízunk abban, hogy az olvasó is egyetért azzal, hogy az angol példák leírásával való megismerkedés nem időpocsékolás, és nem csak az angolul beszélőknek szól, hanem mindenkinek, hiszen ezek példaként szolgálnak arra, hogy milyen publikációk hiányoznak más nyelvek esetében, és még esetleg ötletet is adhatnak egy könyv, segédanyag vagy előadás elkészítéséhez.

5.4.1. Az egynyelvű tanulói szótárakról

5.4.1.1. Bevezetés

A korpusznyelvészet elterjedésében a legnagyobb szerepet véleményünk szerint a nyelvtanulók számára készült korpusz alapú egynyelvű angol szótárak játszották, és

ezek mintegy sűrítve tartalmazzák a korpuszvizsgálatok eredményeit, mindenki számára könnyen hozzáférhető módon. Ezért e fejezetet ezek ismertetésével kezdjük. A szótárak tárgyalásánál óhatatlan az is, hogy ne csak szorosan a korpusznyelvészethez való viszonyukról essen szó, hanem arról is, hogy miért olyan fontos ezek használata a nyelvtanulásban már kezdő szinten is.

Az első korpuszra épülő egynyelvű tanulói szótár a *CollinsCOBUILD English Language Dictionary* (J. Sinclair, 1987a) volt, melynek példáját szinte kivétel nélkül minden szótárkiadó követte. Ez egyben azt is jelentette, hogy a kiadók szótáreladásuk növelése érdekében jelentős ismeretterjesztő szerepet is játszottak, hiszen a termékeiket leíró honlapokon, szórólapokon és előadásokon a korpuszokat és azok használatának előnyeit viszonylag részletesen leírták.

Az egynyelvű szótárak, mint amilyen például a *Magyar értelmező kéziszótár* (Pusztai, 2003) is, azzal a céllal készültek, hogy a nyelv szókincsét összegyűjtsék és pontos jelentésüket vagy jelentéseiket meghatározzák. Az anyanyelvi beszélők számára készült szótárak esetében is több változat készül általában. Vannak teljességre törekvő nagyszótárak és azok rövidített változatai, valamint léteznek gyerekek számára készített szótárak is. Az előbbiek csak terjedelmükben különböznek, az utóbbiaknak viszont figyelembe kell venni a gyerekek nyelvi szintjét és általános ismereteit.

Hasonlóképpen, a nyelvtanulók számára készülő szótáraknak is figyelembe kell venniük a tanulók speciális igényeit. Az angol nyelv tanításában már régen felismerték a lexika és a szótárak szerepét, így az 1920-as évektől kezdve olyan kutatások folytak, amelyek eredményeképpen megszülethetett az első egynyelvű tanulói szótár. A kutatások elsősorban Harold Edward Palmer, Michael West és A. S. Hornby nevéhez fűződtek. Az első ilyen szótár, A *New Method English Dictionary* (West & Endicott) 1935-ben jelent meg, és elsősorban az angol nyelv dekódolását, azaz az olvasást kívánta segíteni. Nem sokkal később, 1942-ben jelent meg az *Idiomatic and Syntactic English Dictionary* (Hornby *et al.*), mely általános használatra készült, és az angol nyelven való kommunikációt is elő kívánta segíteni. Ezt követte 1948-ban egy már ismerősen csengő című, A *Learner's Dictionary of Current English* (Hornby *et al.*). E két publikáció szolgált alapul a ma is jól ismert *Oxford Advanced Learner's Dictionary of Current English* (Wehmeier, 2005) készítéséhez, amelynek legutóbbi, hetedik kiadásában is ott látjuk A. S. Hornby (1898–1978) nevét. Az angol egynyelvű tanulói szótárak története iránt érdeklődők számára feltétlenül ajánljuk az English Dictionaries and Foreign Learners: A history (Cowie, 1999) című könyvet, mely a szótárak elemzése mellett rengeteg érdekes adatot tartalmaz a modern nyelvtanítás korai időszakáról.

5.4.1.2. Anyanyelvi szótár – tanulói szótár

A nyelvtanulók számára készített szótárak nem csak tartalmukban, hanem megírásuk módjában is különböznek az anyanyelvi beszélőknek készültektől. Természetesen mindkét szótár esetében a szótárkészítést anyanyelvi korpusz(ok) és az adott nyelv alapos elemzése előzi meg. Tulajdonképpen az anyanyelvi szótár készítéséhez ez elegendő. A korpusz vizsgálatakor egyes szavak és kifejezések gyakoriságát, kollokációikat, jelentésüket és használatukat elemzik a lexikográfusok. A tanulói szótár esetében nyelvtanulók

munkáiból készült, legtöbbször írott korpuszok elemzésével vizsgálják, hogy a diákok általában milyen hibákat követnek el, hiszen a szótárban valamilyen módon fel kell hívni a figyelmet ezekre a nehezen elsajátítható pontokra. Az idegen nyelv tanulását és tanítását kutató munkák eredményeit és nem utolsósorban a diákok és tanárok igényeit is figyelembe kell venniük a kiadóknak.

Az anyanyelvi korpuszok vizsgálatakor kiderül, hogy adott szavakat milyen gyakran használnak, és ezek a szöveg mekkora százalékát teszik ki. McCarthy (McCarthy, 2004: 11) példaként említi, hogy a Cambridge International Corpus beszélt észak-amerikai angol részében a leggyakrabban előforduló 1800 szó a teljes beszélt nyelvi korpusz 80 százalékát teszi ki. Ebből is kiderül, hogy ha diákjaink a céljaiknak megfelelő szókincset tanulják, akkor gyorsabban és eredményesebben sajátíthatják el az idegen nyelvet. A legtöbb szótárban valamilyen módon jelzik azt, hogy az adott szó mennyire fontos a nyelvtanulás szempontjából (kulcs a Cambridge szótárakban, piros színű szócikk a Longman szótárakban, sarkára állított négyzetek a CollinsCobuild szótárakban stb.). Ilyen információt az anyanyelvi beszélőknek készült szótárak nem tartalmaznak.

A gyakorisági mutatók figyelembevételével a kiadók egy 2–3 ezer szavas szókészletet jelölnek ki, amelyet a szócikkek meghatározásainak megírásához használnak. Így tehát egy szótár 80–100 ezer szócikkét 2–3 ezer szóval is meg lehet magyarázni. Az egynyelvű szótárt sokat használó diák hamar megtanulja azt a szókincset, amit a magyarázatok használnak, hiszen újra és újra ugyanazokkal a kifejezésekkel fog találkozni. A szótár meghatározásainak stílusát elsajátítva könnyen ki tudja majd fejezni magát akkor is, ha a megfelelő szót éppen nem ismeri. Ha a szótár írásakor elegendő 2–3 ezer szó a magyarázatokhoz, akkor a diákok is követhetik ezt a taktikát idegen nyelven való kommunikáció esetén, különösen, ha ez szóban történik, amikor nincs idő arra, hogy szótárban való keresgéléssel töltse valaki az időt. Ezen előnyökön kívül a szótár meghatározásai és példái a nyelvtani szerkezetek elsajátításában is nagy segítségre szolgálnak.

Sok esetben azonban egyes magyarázatok megfogalmazása nehézkes lenne, így az illusztrációk használata egyszerűbb megoldás. A nyelvtanulói szótárakban azonban nem csak ezen esetekben használnak képeket, hanem az idegen nyelven való kommunikáció és a szótanulás elősegítése érdekében tematikus, képes szótárhoz hasonló oldalakat is találunk. Tehát ezekben a szótárakban sokkal több vizuális információ és jelzés kell, hogy segítse a diákot, és e rendszernek jól áttekinthetőnek kell lenni.

Sajnos a magyart idegen nyelvként tanulók számára nem készült még szótár, így magyar példát csak a Magyar értelmező kéziszótárból választhattunk. A szótár CD-ROM változatában a *hód* címszó alatt ezt látjuk:

> Európában kipusztulóban levő, kisebb kutya nagyságú, pikkelyes farkú, rágcsáló vízi állat (Castor fiber): a partba vájt, lakásul haszn. üregek közelében lerágott fadorongokból valóságos várat épít magának.

Az angol anyanyelvi beszélők számára készült szótárban (*The Random House Websters Unabridged Dictionary*, 1997: 184) a *beaver* címszó alatt, mely szintén *hód*ot is jelent, a következőt találjuk:

beaver[1],
Castor canadensis,
Head and body 2½ ft. (0.8 m);
Tail 1 ft. (0.3 m)

bea·ver[1] (bē′vər), *n., pl.* **–vers**, (*esp. collectively*) **–ver** for 1; *v.* –*n.* **1.** a large, amphibious rodent of the genus *Castor*, having sharp incisors, webbed hind feet, and a flattened tail, noted for its ability to dam streams with trees, branches, etc. **2.** the fur of this animal. **3.** a flat, round hat made of beaver fur or similar fabric. 4. a tall, cylindrical hat for men, formerly made of beaver and now of fabric simulating this fur. Cf. **opera hat, silk hat, top hat. 5.** *Informal*. A full beard or a man wearing one. **6.** *Informal*. an exceptionally active or hard-working person. **7.** *Slang (vulgar)*. **a.** a woman's pubic area. **b.** *Offensive*. a woman. **8.** *Textiles*. **a.** a cotton cloth with a thick nap, used chiefly int he manufacture of work cloethes. **b.** (formerly) a heavy, soft, woolen cloth with a thich nap, made to resemble beaver fur. **9.** (*cap.*) a native or inhabitant of Oregon, the Beaver State (used as a nickname). – *v.i.* **10.** *Brit.* to work very hard or industriously at something (usually fol. by *away*). [bef. 1000; ME *bever*, OE *beofor*, *befor*; c. G *Biber*, Lith *bebrùs*, L *fiber*, Skt *babhrús*] reddish brown, large ichneumon] – **bea′ver·like, bea′ver·ish**, *adj*.

beav·er[2] (bē′vər), *n*. *Armor*. **1.** a piece of plate armor For covering the lower part of the face and throat, worn Esp. With anopen helmet, as a sallet or basinet. Cf. **buffe, wrapper** (def. 7) **2.** a piece of plate armor, pivoted atthe sides, forming part of aclose helmet below the visor or ventail. See diag. under **close helmet**. [1400-50; late ME *bavier, bavour* < MF *baviere* OF: bib), equiv. to *bave* spit, dribble + *-iere* < L *-āria*, fem. of *-āruis* –ARY; alteration of vowel int he initial syll. is unexplained]

55. ábra: A *beaver* címszó az anyanyelvi szótárban

A fentiekből láthatjuk, hogy a főnévi (9 db) és igei jelentések (1 db) egy szócikken belül szerepelnek, melyek között a vulgáris is megtalálható (7-es szám). Jelen esetben a számunkra legérdekesebb a legelső pont. Ha ezt megnézzük, akkor valószínű, hogy az angolul jól tudó olvasók is találnak számukra ismeretlen szót vagy szavakat. A szócikk végén szögletes zárójelben a szó etimológiáját is megtaláljuk, amire valószínű, hogy egy nyelvtanulónak semmi szüksége nincs.

A gyermekek számára készült szótárban talán kicsit érthetőbb meghatározást találunk (*Merriam-Webster Dictionary for Kids Online* http://www.wordcentral.com/):

Main Entry: **[1]bea·ver**
Pronunciation: 'bE-v&r
Function: *noun*
Inflected Form(s): *plural* **beaver** *or* **beavers**
1 : a large fur-bearing mammal that is related to the rats and mice, has webbed hind feet and a broad flat tail, and builds dams and underwater houses of mud and branches
2 : the fur of a beaver

56. ábra: A *beaver*[1] címszó az anyanyelvi gyermekszótárban

Az egyes pontban levő meghatározást látva is találunk olyan szavakat (pl. *mammal* és *hind*), amit egy magyar anyanyelvű nyelvtanuló valószínű, hogy nem értene. Érdekes, hogy kép nem szerepel a meghatározás mellett, csak az itt nem feltüntetett jelentése esetében, mely középkori páncélsisak egy bizonyos részére vonatkozik. Azt is megfigyelhetjük, hogy az előbbi 9 főnévi jelentésből itt mindössze kettő szerepel és az igei használat is külön szócikket képez.

A Merriam-Webster online, azaz frissebb változata sem jobb az első meghatározásnál:

[1] bea·ver 1 *or plural* **beaver a:** either of two large semiaquatic herbivorous rodents" (*Castor canadensis* of No. America and *C. fiber* of Eurasia) having webbed hind feet and a broad flat scaly tail and constructing dams and partially submerged lodges

Nézzünk meg, hogy a tanulói szótárak hogy határozzák meg ugyanezt a szót.

bea·ver[1] /ˈbi:və $ -ər/ *n* [C] a North American animal that has thick fur and a wide flat tail, and cuts down trees with its teeth → **eager beaver** at **eager (2)**
beaver[2] *v*
beaver away *phr v informal* to work very hard, especially at writing or calculating something: [+**at**] *He's been beavering away at his homework for hours.*

57. ábra: A *beaver* meghatározása a *Longman Dictionary of Contemporary English* szerint

bea·ver /ˈbi:və(r)/ *noun, verb*
- *noun* **1** [C] an animal with a wide flat tail and strong teeth. Beavers live in water and on land and can build DAMS (= barriers across rivers), made of pieces of wood And mud. – see also EAGER BEAVER – picture on page A6
  2 [U] the fur of the beaver, used in making hats and clothes
- *verb* PHR V, **beaver aˈway (at sth)** *(informal)* to work very hard at sth: *He's been beavering away at the accounts all morning.*

58. ábra: A *beaver* meghatározása az *Oxford Advanced Learner's Dictionary* szerint

bea|ver /biːvər/ **(beavers, beavering, beavered)**
1 A **beaver** is a furry animal with a big flat tail and large teeth. Beavers use their teeth to cut wood and build dams in rivers. **2 Beaver** is the fur of a beaver. ❑ *…a coat with a huge beaver collar.*
♦ **beaver away** If you **are beavering away at** something, you are working very hard at it. ❑ *They had a team of architects beavering away at a scheme for the rehabilitation of District 6… They are beavering away to get everything ready for us.*

59. ábra: A *beaver* meghatározása a *CollinsCOBUILD* szerint

A fenti példák jól szemléltetik, hogy annak ellenére, hogy sok tekintetben más elrendezésben szerepel az információ, mindegyik hasonlít abban, hogy csak a legszükségesebb és legvalószínűbben előforduló jelentéseket sorolják fel. Azt is észre vehetjük első pillantásra, hogy a *CollinsCOBUILD* szótárnál a nyelvtanra vonatkozó információt külön oszlopban találjuk a jobb oldalon, és a meghatározásban teljes mondatok szerepelnek, hogy ezzel is segítsék a diákokat a helyes használat elsajátításában.

A tanulói szótárak manapság olyan módon készülnek, hogy akár tankönyvként is használhatók. A nyelvtanuláshoz szükséges szinte minden információt megtalálhatunk bennük. Mivel korpuszra épülnek, a nyelv változását is jól követik, és nem elavult információval szolgálnak. Sok esetben gyakorló oldalak (study pages) találhatók a szótárban, melyek a szótárhasználat elsajátítását és ezzel egy időben bizonyos nyelvtani problémák gyakorlását is lehetővé teszik. Alapvető nyelvtani tudnivalók, szóhasználati tanácsok, hibás és helyes használatra való figyelmeztetések, levélminták és sok egyéb hasznos tanács található a különböző szótárakban a szavak jelentése mellett. A korpuszból származó példamondatokról sem szabad megfeledkeznünk, amelyek mintául szolgálnak a diákok számára, valamint az egyéb korpusz alapú információkról sem. Ilyenek lehetnek például az adott szó leggyakoribb kollokációi: ***make/reach*** *a decision*; *a* ***difficult / final/ important/ unanimous / wise*** *decision*; *a decision* ***about / on*** *sth* (*Cambridge learner's dictionary (Semi-bilingual version)*, 2004: 207), vagy a *hate* szócikknél grafikon szemlélteti, hogy a *hate* igét majdnem 80%-ban valaki vagy valami követi (pl. *I hate English*), körülbelül az esetek 10–10%-ában használják a *hate doing sth* vagy a *hate to do sth* szerkezetet, melyet a *hate it when* követ, és az ezeken kívüli más szerkezetek elenyésző számban fordulnak elő (*LDOCE*, 2003: 744). A szótárak nagy kínálata ellenére kevés azonban a kollokációs szótár. A Magyarországon is kapható *Oxford Collocations Dictionary for Students of English* (Lea, 2002) a 100 millió szóból álló Brit Nemzeti Korpusz elemzésének eredményeit használta fel a szótár készítéséhez. A magas szintű nyelvtudásra törekvő diákok számára elengedhetetlen segédeszköz egy ilyen szótár.

Az egynyelvű szótárakat a legtöbbször azzal hárítják el a diákok, és sokszor a nyelvtanárok is, hogy legalább középfokú vagy magasabb nyelvtudásra van szükség a használatukhoz. Az első tanulói egynyelvű szótárak valóban a középfokon álló diákok számára készültek, de ma már számos szótár közül választhatunk a tudásszintnek és korosztálynak megfelelően. A nyelvkönyvet sem úgy vesszük, hogy egész életünkben azt az egyet

használjuk, így a szótárakat is „kinőjük”. A kezdőknek tehát az ő számukra írt, kevesebb szócikket és jelentést felsoroló szótárra van szükségük kezdetben, majd tudásuk gyarapodásának megfelelően kell később más szótárt választani. A legtöbb iskolában és tanfolyamon nem fordítanak sok figyelmet a szótárhasználat tanítására és gyakorlására, pedig a szótárak gyors és eredményes használata egyik fő feltétele az önálló tanulásnak is. Azt is láthattuk az eddig leírtakból, hogy az egynyelvű szótárak sok olyan információval szolgálnak, amelyet a kétnyelvű szótárakban nem találhatunk meg.

A tanulói egynyelvű szótárak szinte kivétel nélkül CD-ROM-on is megjelentek, és a papíron kiadottal együtt megvehetők, így tehát egy kis árdifferenciával egyszerre két formátumban is hozzájuthatunk. Így például a *MacMillan English Dictionary for Advanced Learners* CD-ROM melléklettel 5438 Ft és anélkül 4980 Ft volt 2004 decemberében. A CD-ROM jellegénél fogva olyan keresési módokat és interaktív gyakorlatokat is tartalmaz, amelyet a hagyományos könyv formájában nem lehet megoldani. Így az elektronikus változatban a keresés eredménye egy témakörök szerinti teljes szólista is lehet, amely az idegen nyelven való fogalmazást nagyon megkönnyítheti. Ezen kívül a példamondatok nem csak az adott szócikkből, hanem a teljes szótárból megjeleníthetők, így az egy vagy két példa helyett akár tízet is megnézhet a diák. A modern szótárak készítői a tanárokra is gondoltak, így például a *CollinsCOBUILD* szótár CD-ROM-ja egy 5 millió szavas mini korpuszt is tartalmaz, az új kiadású *LDOCE* CD-ROM-ján pedig tanítási segédanyagokat is találunk.

5.4.2. COBUILD kiadványok

5.4.2.1. Tankönyv

Az első korpuszra épülő tankönyv nem sokkal az első szótár, a *Collins COBUILD* szótár megjelenése után került a piacra *Collins Cobuild English Course* (J. R. Willis & Willis, 1988) címmel. A tankönyv teljes egészében a korpuszvizsgálatok eredményeire épült, és a szókincset is a gyakoriság alapján választották ki. Ekkor ezt az elvet azonban túlságosan is szigorúan követték, így fordulhatott elő, hogy a különböző színek különböző helyen szerepeltek a tananyagban. Minden szemantikailag összetartozó csoportban vannak gyakrabban és ritkábban használt szavak vagy kifejezések, de a nyelvtanulás eredményessége érdekében célszerű ezeket együtt megtanulni, könnyebb ezekre így emlékezni. Valószínű azonban, hogy a könyv elterjedésének legfőbb akadálya az volt, hogy túlságosan is megelőzte a korát. Erre talán az is bizonyíték, hogy mint látni fogjuk, a következő hasonló tankönyv csak idén, több mint 15 évvel később jelent meg.

5.4.2.2. Segédanyagok

A COBUILD projekt ismertetésekor már említettük, hogy a két fő cél közül az egyik az volt, hogy a korpuszelemzések eredményeit az angol nyelvet tanuló diákok és oktatók számára készülő referencia és tankönyvek írásakor felhasználják és publikálják (Krishnamurthy, 1997b). Az 1990-es évek elején a számítástechnika még nem állt olyan szinten, mint manapság, így a korpuszokhoz való hozzáférés elsősorban a helyszínen

történhetett meg. Ma az előző fejezetben bemutatott konkordanciaprogramok és az internet vagy CD-ROM-ok felhasználásával bárki könnyedén készíthet és kinyomtathat konkordanciákat, ami abban az időben szinte elképzelhetetlen volt. Ezért tartották fontosnak, hogy konkordancia gyűjtemények kiadásával segítsék a tanárok munkáját. A *Collins Cobuild Concordance Samplers*nek négy kötete jelent meg: 1. *Prepositions* (Capel, 1993); 2. *Phrasal Verbs* (Goodale, 1995a); 3. *Reporting* (G. Thompson, 1995); és 4. *Tenses* (Goodale, 1995b). Ezek a kötetek annyira feledésbe merültek, hogy még a kiadó munkatársai sem igen hallottak róluk, és már csak antikváriumban akadhat rájuk az ember véletlenül. Arra azonban most is jó példaként szolgálhatnának, hogy milyen konkordanciákat érdemes választani a diákokkal való munkához.

A *Collins COBUILD English Guides* sorozat is a korpuszelemzések eredményeként jött létre, és elsősorban olyan nyelvtani pontokat kívánt bemutatni, amelyek a nyelvtanulóknak nehézségeket okoznak. A tíz kötetből álló sorozat első része 1991-ben, az utolsó pedig 1997-ben jelent meg: 1. *Prepositions* (J. Sinclair & Cobuild, 1991a); 2. *Word Formation* (J. Sinclair & Cobuild, 1991b); 3. *Articles* (Berry & Cobuild, 1993); 4. *Confusable Words* (Carpenter & Cobuild, 1993); 5. *Reporting* (G. Thompson & Cobuild, 1994); 6. *Homophones* (J. Sinclair & Cobuild, 1995); 7. *Metaphor* (Deignan & Cobuild, 1995); 8. *Spelling* (J. A. Payne *et al.*, 1995); 9. *Linking Words* (Chalker & Cobuild, 1996); és 10. *Determiners & Quantifiers* (Berry & Cobuild, 1997).[30] Egyes kötetek gyakorlatokat is tartalmaznak a leírás mellett.

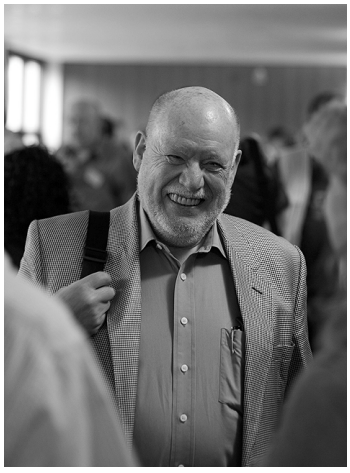
Természetesen hagyományosabb jellegű segédanyagok is készültek, mint például a különböző méretű és mélységű nyelvtanok: *Basic Grammar* (D. Willis & Wright, 1995), *Student's Grammar* (Watson *et al.*, 1991) és az *English Grammar* (J. Sinclair, 1990). Úttörő jellege miatt ez utóbbit ki kell emelnünk. Szinte megszámlálhatatlan az angol nyelvtanok száma, de ez a leíró nyelvtan az első olyan átfogó mű, amely teljes egészében korpuszvezérelt. Ez azt jelenti, hogy a nyelvtani leírás a korpuszadatok tényleges elemzésének eredményeit összegzi, és nem pusztán nyelvtani pontok illusztrálásának céljából, vagy hipotézisek ellenőrzésére használták. Az elemzéshez használt korpusz több mint 100 millió szóból állt.

A hagyományos nyelvtanok többsége elsősorban az írott nyelvre vonatkozik, a *Collins COBUILD English Grammar* viszont a beszélt és az írott nyelv közötti nyelvhasználati különbségeket is figyelembe veszi, és felhívja ezekre a figyelmet. Célja, hogy teljes képet adjon az angol nyelv különböző reprezentációs szintjeiről a szavaktól a diskurzus szintjéig. A könyv szerkezetét ismertető diagram a xxiv-xxv oldalon található. A nyelvtanulók és nyelvtanárok számára egyaránt hasznos kézikönyv további előnye még az is, hogy a gyakori nyelvtani szerkezetek részletes leírása mellett az egyes szerkezetekben leggyakrabban előforduló szavak listája is megtalálható, valamint a beszélő szándékára vonatkozó magyarázatok. Olyan esetekben, ahol nagy a diákok tévedésének valószínűsége, külön figyelmeztetést találunk a nyelvtani leírás mellett. A produktivitás feltűntetése a diákok figyelmét olyan jelenségek tanulására irányítja, amelyek „kifizetődőbbek", hiszen gyakrabban fordulnak elő. A fejezetek címeit más

[30] A kötetek címei magyarul: 1. Elöljárószók; 2. Szóképzés; 3. Névelők; 4. Összetéveszthető szavak; 5. Hírül adás; 6. Azonos hangzású szavak; 7. Metaforák; 8. Helyesírás; 9. Kötőszavak, 10. Determinánsok és kvantorok.

nyelvtanokéval összehasonlítva is jól láthatjuk a szemléletbeli különbségeket. Míg a legtöbb nyelvtani leírás szófajok és nyelvtani fogalmak neveit választja fejezetcímként, itt ismét a nyelvhasználat kerül előtérbe, hiszen olyan fejezetcímeket találunk, mint pl. „1. Fejezet: Emberekre és tárgyakra való utalás", vagy „2. Fejezet: Emberekre és tárgyakra vontakozó információk kifejezése".

A nyelvtan feldolgozásakor olyan korpuszelemzésre épülő összefoglaló művek is készültek mint a *Grammar Patterns 1: Verbs* (J. Sinclair, 1996b) és a *Grammar Patterns 2: Nouns and Adjectives* (J. Sinclair, 1998). Ezek a könyvek abban segítenek, hogy elsajátítsuk bizonyos igék, főnevek, és melléknevek használatát, amelyek tipikus szerkezetekben fordulnak elő. Például a főnevek jelentős részét tetszés szerint használhatjuk jelzős szerkezetekben vagy önmagukban. Vannak azonban olyanok is, amelyek vagy kizárólagosan vagy tipikusan csak jelzővel együtt fordulnak elő. Ilyenek például a *creature, listener, margin, soil* és a *wing* (J. Sinclair, 1998: 80). Ezek a segédkönyvek még az angol anyanyelvű tanár könyvespolcáról sem hiányozhatnak. További nyelvhasználati és szókincsfejlesztést elősegítő könyveik a *Vocabulary Builders* 1–4, (*Keywords in the Media, Keywords in Business* stb.), valamint az egynyelvű tanulói szótárak mellett más speciális szótárakat is kiadtak, amelyek közül csak néhányat említünk meg példaként: idióma, úgynevezett „phrasal verbs"[31] szótárak, amelyekhez általában munkafüzetet is készítettek. Természetesen az 1990-es években kiadott könyveket és szótárakat egyfolytában javítják, újabb kiadások jelennek meg. A *Collins Cobuild Student's Dictionary* harmadik kiadása 2005-ben jelent meg.



60. ábra: John Sinclair az ICAME 25 Konferencián Veronában
(Sebastian Hoffmann képe, http://es-sebhoff.unizh.ch/pictures/photo.php)

E könyv nem lehetne teljes anélkül, hogy ne említsük meg külön kiemelve John Sinclairt, aki az empirikus nyelvszemlélet egyik legmarkánsabb képviselője, és a korpusznyelvészet egyik megteremtője. Neve egybeforrt a COBUILD projekttel és a Birminghami Egyetemmel. Szinte egyetlen COBUILD publikáció sem készült aktív

[31] Igéből és határozószói/elöljárószói partikulából álló szerkezet, pl. *look for, look after, take after* stb.

közreműködése nélkül, így e könyvben is számos referencia található nevével. Nem csak nyelvészként, hanem pedagógusként is nagy hatással volt a nyelvészet és a nyelvoktatás fejlődésére. John Sinclairnek köszönhető, hogy az elmúlt 15 évben egyre több és jobb korpuszra épülő vagy azt felhasználó segédkönyv és tankönyv áll a nyelvtanulók és nyelvtanárok rendelkezésére, hiszen a COBUILD projekt úttörő publikációit más kiadók is követték egymással versenyre kelve.

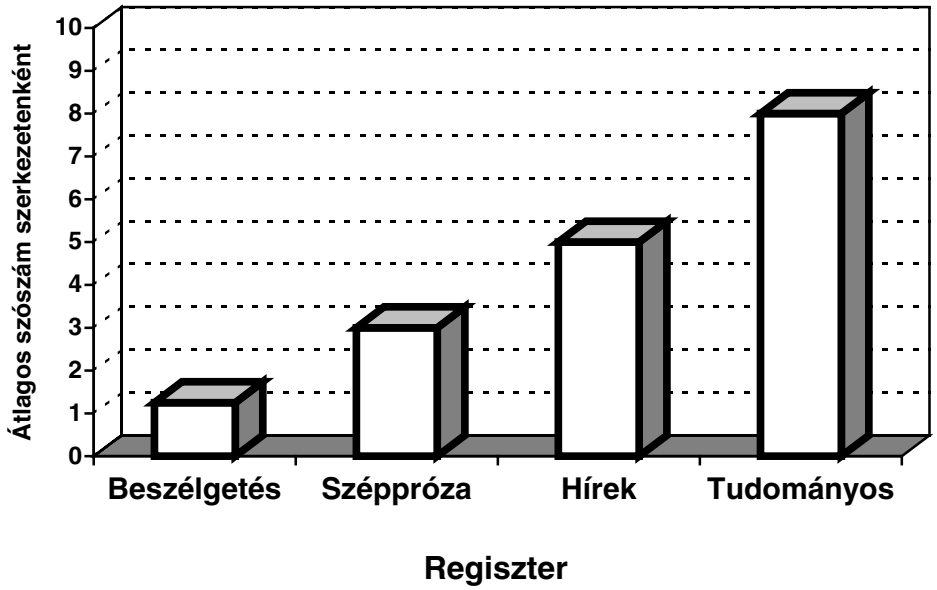
5.4.3. A Longman Grammar of Spoken and Written English (LGSWE)

David Crystal úgy nyilatkozott az LGSWE (Biber *et al.*, 1999) könyv láttán, hogy ha még aktívan tanítaná az angol nyelvtant, akkor hajlana arra, hogy összes jegyzetét eldobja és elölről kezdjen mindent (Longman szórólap IPN: 0997 023228). Már a cím is sokatmondó, hiszen a legtöbb nyelvtan az írott nyelv normáit írja le és a beszélt nyelv ennek fényében legtöbbször csak „nyelvtanilag helytelen"-ként kerül említésre. Harold Palmer volt az egyetlen, aki e könyv megjelenése előtt a beszélt nyelv nyelvtanát vizsgálta, aminek eredményét 1924-ben megjelent könyvében összegezte (Palmer, 1924). Az LGSWE szisztematikusan vizsgálja a beszélt nyelvet úgy, hogy nem előre meghatározott szabályok alapján akarja eldönteni, hogy helyes-e, hanem egyszerűen leírja, hogy a beszélt nyelv nyelvtana valójában milyen is. E nyelvtan segítségével az is lehetővé válik, hogy a nyelvet idegen nyelvként tanító tanárok és tanuló diákok a beszélt nyelvet is jobban megismerhessék a nyelvterületen való élés vagy anyanyelvi beszélőkkel való kapcsolat nélkül is, ami eddig szinte lehetetlen volt.

Természetesen az írott nyelv vizsgálatát a beszélt nyelvhez hasonlóan úgy közelítik meg, hogy nem példákat keresnek nyelvtani szabályok illusztrálása céljából, hanem a korpuszadatok elemzésére alapozva írják le, hogy valójában hogyan is használják a nyelvet. Talán az újdonság bizonyítéka az is, hogy az első 45 oldalt az angol nyelvtan korpusz alapú megközelítésének leírásának szentelték, amelyben részletesen leírják céljaikat, módszereiket.

Mire is épül az LGSWE elemzése? Egy több mint 40 millió szóból álló korpuszra, amelynek legfőbb elemei brit beszélgetések (közel 4 millió szó), brit és amerikai szépprózа (közel 5 millió szó), brit hírek (kb. 5,4 millió szó), brit és amerikai tudományos próza (kb. 5,3 millió szó). A dialektusok összehasonlítása céljából szerepel benne majdnem 2,5 millió szóból álló amerikai beszélgetés és kb. 5,2 millió szóból álló amerikai hírek gyűjteménye. Ezt egészíti ki a nem beszélgetést tartalmazó brit beszéd (kb. 5,7 millió szó) és a brit és amerikai általános próza (kb. 6,9 millió szó) (Biber et al., 1999: 25).

Az LGSWE számos összehasonlító táblázatot és ábrát tartalmaz, amelyekre sokszor elég csak egy pillantást vetni, és máris kíváncsian olvassuk a hozzájuk fűzött elemzést vagy magyarázatot. E stílus érzékeltetésére egy Geoffrey Leech által készített ábrát mutatunk be, amit ugyan nem az LGSWE-ből vettünk, de az adatok ugyanarra a korpuszra vonatkoznak (Leech, 1998). A szószerkezetre egy példa: *a sarokban ülő fiú*, mely négy szóból áll. Az ábrán azt láthatjuk, hogy a beszélgetések esetében sokkal rövidebb szerkezeteket találhatunk, mint írott szövegekben.



61. ábra: Átlagos szószám szerkezetenként a különböző regiszterekben

Mint azt az LGSWE címe is elárulja, elsősorban a beszélt és írott nyelv nyelvtani különbségeire szándékszik fényt vetni, de a felsorolt alkorpuszok is mutatják, hogy a brit és amerikai nyelvhasználat közötti különbségeket is vizsgálták, még ha ez kisebb szerepet is kapott. A könyvnek készült egy diákok számára írt változata (Biber *et al.*, 2002), melyhez munkafüzet is kapható.

5.4.4. Touchstone – új korpusz alapú tankönyv

Jane és Dave Willis korpuszra épülő tankönyve (1988) után ez a második ilyen jellegű tankönyv. Mint említettük, a Willis házaspár könyve talán túlságosan korán jelent meg, így nem lett népszerű. A *Touchstone* (McCarthy *et al.*, 2005) éppen, hogy csak megjelent, ezért nehéz megjósolni, hogy milyen sikere lesz, hány tanár választja majd iskolai tankönyvként. Annyit azonban máris megállapíthatunk, hogy a nyelvtanári közösség Japánban, ahol a könyvet először bemutatták, nagy érdeklődést mutatott. A tankönyv korpuszra épül, de a hagyományos tankönyvírás elvei is megmutatkoznak benne. A gyakoriság nem elsődleges szempont a tankönyvbe kerülő szavak és szerkezetek kiválasztásakor, így a kevésbé gyakori, de az eddigi tapasztalat szerint szükséges szerkezetek és szókincs is megfelelő arányban szerepel.

A kapcsolatteremtés idegen nyelven még nehezebb, mint az anyanyelven, de a nyelv és a nyelvhasználat jobb megismerésével tudatosan is fejleszthető. A kutatások alapján megállapítható, hogy a beszélt nyelvben leggyakrabban előforduló 50 szó jelentős része az „én” és „te”, azaz a beszélő és hallgató közötti kapcsolatteremtést és nem kimondottan az információközlést szolgálja. Így például az egyik leggyakoribb kifejezés a *you know* (tudod), amely a beszélő és hallgató által is ismert dolgokra való utalással erősíti

meg a közöttük levő kapcsolatot. A *really* vagy *that's right* kifejezések az „aktív hallgatás" kellékei, melyek szintén a beszélővel való együttérzést vagy együttértést fejezhetik ki. A beszélt nyelvben fontosak az egy szuszra kimondott egységek, mint például a magyarban a „*ne haragudjon, hogy zavarom, de*", így az ilyen jellegű lexikai csoportok begyakorlását is segítik a párbeszédek. A tankönyv tehát az elemzések eredményeit felhasználva olyan szófordulatok elsajátítását segíti a korpuszból származó autentikus párbeszédek felhasználásával, amelyek megkönnyítik a nyelvtanulók idegen nyelven való kapcsolatteremtését.

5.5. Saját készítésű feladatok

Az eddigiekben azt néztük meg, hogy a korpusznyelvészet milyen „késztermékei" állnak rendelkezésünkre a nyelvtanításban. Nem véletlen, hogy csak angol nyelvre vonatkozó példákat hoztunk, hiszen másról, sajnos, egyelőre nem tudunk beszámolni információ hiányában, ami a nem létezésükre utal. Így a magyar vagy idegen nyelvet tanítóknak és tanulóknak a következőkben igyekszünk tanácsot adni, hogy hogyan használhatják a korpusz módszereket a tanulásban és tanításban.

Véleményünk szerint az első lépés a tanítás felé az, hogy a tanár maga is használja a korpuszt bizonyos kérdések megválaszolására vagy egyszerűen vizsgálódás céljára. Ehhez a legmegfelelőbb az interneten elérhető korpuszok használata, hiszen ezekhez keresőprogramok is tartoznak, így könnyű használni őket. Az ilyen korpuszokat sokszor annotálták is, ami tovább könnyíti a gyors és pontos információ lekérdezést. Az a tény sem hanyagolható el, hogy az ilyen honlapokon alapos súgókat is találunk, és a „gyakran feltett kérdések" rovat is segíthet, ha elakadunk.

A második lépés az, hogy korpusz alapú kinyomtatott vagy számítógépes gyakorlat készítéséhez használja fel a tanár a korpusz adatait. A legideálisabb esetben a diák maga végzi a „kutató" munkáját, és a korpuszt saját maga vizsgálja. Ehhez olyan korpuszt kell összeállítani, amelyet a diákok önállóan képesek használni. Tim Johns találó kifejezést használt erre: adat-vezérelt nyelvtanulás (data-driven language learning, rövidítve DDL), melynek honlapját a következő címen lehet elérni: http://web.archive.org/web/20040203111227/http://web.bham.ac.uk/johnstf/timconc.htm.

A címből kitűnik, hogy ez egy archívumban szerepel. Nem tudjuk, hogy meddig lehet majd ott megtalálni, így javasoljuk, hogy az érdeklődők mentsék el saját gépükre az információt.

5.5.1. Konkordanciák nyomtatásban

A kinyomtatott konkordanciák lehetőséget adnak arra, hogy a tanár saját maga válassza ki a keresett szó vagy kifejezés azon előfordulásait az anyanyelvi korpuszból, amelyek diákjai nyelvi szintjének vagy a tanítandó elemek szempontjából a legmegfelelőbbek. Mivel a hagyományos módon készített konkordanciák általában nem teljes mondatokban láthatók a képernyőn, a konkordanciákat érdemes az első néhány alkalommal teljes mondatok formájában használni, de a kulcsszó középen való elhelyezését megtartani.

Az ilyen jellegű konkordanciák alkalmasak a „tudatosság növelésére”, mivel a diákokat a megfigyelésre és következtetések levonására ösztönzi. A megfigyelés szempontjai vagy a választott mondatok lehetnek nagyon egyszerűek, így akár a legelső óráktól kezdve használhatók az idegen nyelv és az anyanyelv tanítására is.

Nyelvtani jellegű gyakorlatok

A nyelvtan tanítása nem egyszerű még az anyanyelv esetében sem, és általában nem tartozik a diákok kedvencei közé. A konkordanciák segítségével azonban játékosabb formában lehet a *szófajok felismerését* és megkülönböztetését gyakorolni. A korpusz és a konkordanciák használatának legfőbb előnye az, hogy egy-egy jelenségre sokkal több példát tudunk rövid idő alatt produkálni, mint ha magunk próbálnánk listákat vagy példamondatokat készíteni. Így a diákoknak bőséges példaanyag áll rendelkezésükre. A *vár* szóra keresve az MNSZ-ben, percek alatt a 12 722 példából 500 darabot tudunk megnézni és kimásolni. A keresett szó mellett áll szófajra vonatkozó információ is, amit ki is törölhetünk, ha nem tartjuk szükségesnek. Ha azonban azt szeretnénk, hogy a diákok később maguk is használják a korpuszt, akkor érdemes ezt megtartani, és apránként megismertetni velük a kódokat.

1.	ideig. Azonban később a	**vár**	*N.NOM*	őrségét kiéheztették, mire azok
2.	edénytöredékek sorában említhetjük pl. a	**vár**	*N.NOM*	kőépületének belsejéből napfényre bukkant,
3.	nem tudták befejezni. A	**vár**	*N.NOM*	így is állta a támadásokat
4.	a besúgó aki csak arra	**vár**	*V.e3*	, hogy ezúton érdemeket szerezzen
5.	- 1998._december_12., szombat Börtön	**vár**	*V.e3*	a román hírszerző tisztre Az
6.	Ölvedi Ignác: A budai	**vár**	*N.NOM*	és a debreceni csata (
7.	tanúk tanúi legyenek a büszke	**vár**	*N.NOM*	megaláztatásának is. Lakodalom sem
8.	alapterületű emeleti szintjén az egri	**vár**	*N.NOM*	történetével. Deák Endre és
9.	fordulat előtt. Amerika folytonosságot	**vár**	*V.e3*	Bármi lesz is a választások
10.	: a világra gazdasági hanyatlás	**vár**	*V.e3*	, ha nem reformálják meg

62. ábra: A *vár* konkordanciája az MNSZ-ből

A nyelvtani gyakorlatok egy másik változatánál az **egyeztetést** lehet gyakorolni. A következő angol nyelvű feladatban az egyes számban levő igét (*is*) olyan főnévi szerkezet előzi meg, mely többes számú főnévvel végződik. Azt kell a diáknak megmondania, hogy melyik az a főnév, amellyel az igét egyeztetjük.

```
1) December could not agree. So an eight man team of scientists is to
mae a lengthy tour of
2) sists of all-carbon rings. In a few rings one of the carbons is
replaced by an atom of oxy
3) Britain's only major industry in digging up metals is based in
Cornwall. There a
```

```
 4) petitive, not monopolistic. Each of the five serious dailies is
separately owned and edite
 5) der to reduce noise, The choice of species in tree plantings is
also important. Broad leav
```

63. ábra: Gyakorlat az egyeztetésre (Tim Johns)

A megfelelő példák kiválasztásával bármilyen nyelvtani jelenséget lehet gyakoroltatni, ugyanúgy, mint a hagyományos feladatkészítések során. A különbség annyi, hogy itt autentikus, anyanyelvi beszélők által használt, kész mondatokkal dolgozunk, és nem a tanárnak kell kitalálnia a példákat.

Íme egy példa a névelők használatának gyakorlására. Mint a konkordanciákból jól látható, itt nem egyszerűen a kulcsszót töröltük ki, hanem az előtte szereplő névelőt is. A kulcsszó névelős és névelő nélküli használatát szemléltető példamondatok alapos megfigyelése után, tehát valamilyen szabály megfogalmazásával tudja a diák ezt a feladatot megoldani.

trade

1. The arms trade brings misery to millions in diverting spending from social needs to military ones, . . .	2. Most supermarket chains are not yet interested in taking fair trade further than one brand of coffee or chocolate, . . .

```
 1. vegetation. At the height of _____ slave trade, Cape Verde
provided a transit camp for the s
 2. r hope to feed, that _____ international trade is carried out
unfairly and unequally. But wi
 3. s - a man deeply involved with _____ drugs trade and the death
squads. Then there were the t
 4. iated the restrictions on _____ Japanese trade. In March 1979
Haferkamp visited Tokyo to dis
 5. atives of the pet industry and _____ fur trade, with both factions
desperately trying to sel
 6. rkers and drivers." In _____ film and TV trade HRH is known as
"One take Charlie". His polis
 7.  breach of the EEC's rules on _____ free trade. The Commission's
verdict threatened to sweep
 8. The ancient fraternity of _____ printing trade, with its complex
hierarchy of "chapels" (unp
 9.  successes  -  _____  international  wildlife  trade,  worth  an
incredible ⌋1000 million a year, ta
10. in more about the history of _____ local trade and perhaps show
how knowledge of bronze cast
11. unnecessary regulations on _____ nuclear trade. Other nations,
however, are carping about a
12. f the first millennium BC. _____ incense trade continued to be
important for the next 2500 y
```

64. ábra: Gyakorlat a névelők használatára (Tim Johns)

Lexikai jellegű gyakorlatok

A nyelvtanulásban kinek a nyelvtan tűnik nehéznek, ki meg a szavak végeláthatatlan sorát tartja megtanulhatatlannak. A nyelvtant jól használó, nagy szókinccsel rendelkező beszélők esetében is a legtöbb problémát azok a kifejezések okozzák, amelyeket nem lehet úgy megtanulni, mint az idiómákat. Ezek azok a kifejezések, amelyek esetében bizonyos szabadságot élvezünk a szavak megválasztásában, de ez a szabadság limitált. A szótárakban általában nem kapunk elég információt ezekről, így igazából más forrásokból kellett ezeket „összeszedni". Talán nem sokan használják, de léteznek kollokációs szótárak, melyek közül a legkönnyebben beszerezhető a *BBI,* azaz *The BBI Combinatory Dictionary of English: A guide to word combinations* (Benson *et al.*, 1986). Ha a kollokációkat meg is értjük, idegen nyelven való fogalmazás során szembesülünk a problémával, és ebben segít e szótár használata is. Éppen a nehézségük miatt, a konkordanciák segítsége talán a kollokációk terén a legfontosabb.

Kollokációk vizsgálata

A 65. ábra keresési eredményeit többféle feladathoz lehet felhasználni külön-külön és együtt is. Alsó tagozatos magyar anyanyelvű gyerekek esetében egyszerűen megkérdezhetjük, hogy szerintük van-e különbség a *piros* és a *vörös* között. Ha igen, akkor mi a különbség? Már ezzel az egyszerű kérdéssel is arra hívjuk fel a figyelmet, hogy sokszor a felületesen „majdnem egyforma" kifejezések is különbségeket takarnak. Második lépésként készítsenek a gyerekek egy listát arról, hogy mi *piros* szerintük és mi *vörös*, azaz milyen szavakat használunk ezekkel. Bizonyára sok játék neve is előfordul majd, mint például *piros labda*, *piros tűzoltóautó*, de a *vörös bor*, vagy a *vörös zászló* is talán elő fog fordulni. A konkordancialistákat csak ezek után használjuk. Célszerű minél több, de nem ábécé sorrendbe rendezett példát adni. A 65. és 66. ábra ábécérendben mutatja a példákat, így azonnal látható, hogy a *vörös ördögök* vagy a *piros alma* kétszer is szerepel. Ez így túl egyszerű feladat lenne, hiszen csak a sorokat kellene megszámolni. Ezt a megoldást is választhatjuk, de ha a példamondatok az előfordulás szerint szerepelnek, azaz nem ábécé rendben, akkor a gyerekeknek a látott információt rendezni is kell, és ez a tanulást és a memóriát jobban segíti. Például listát írhatnak az előfordult kifejezésekről és mellette számmal jelezhetik az előfordulások számát. A diákokat két csoportra osztva, egyik csoport a *piros*, a másik a *vörös* előfordulásait tanulmányozhatja. Következő lépésként rákérdezhetünk, hogy hány olyan kifejezést találnak, amikor a valóságos színt jelenti a szó, és hány esetben használják valami más értelemben. Színes ceruzák használatával egyszerűen két különböző színnel jelezhetik ezt (pl. *Vörös Hadsereg*, *vörös báró*). Ha szükséges, a tanár kérdésekkel vezetheti rá a diákokat a helyes válaszra. A csoport megfigyeléseit ezek után először ajánlott a tanár vezetésével megbeszélni. Ha a diákok rendszeresen végeznek ehhez hasonló feladatokat, akkor a későbbiekben kis csoportokban vagy párokban is megbeszélhetik „felfedezéseiket". Természetesen a konkordanciák vizsgálata előtt írt listákkal is össze lehet hasonlítani az eredményt. Hány olyan szókapcsolat szerepel a konkordanciákban, amit nem írtak le a gyerekek, vagy fordítva. A *piros* és *vörös* felcserélhetőségét is megvizsgálhatjuk a példák alapján. Mondjuk-e vajon azt, hogy *piros bor* vagy *vörös autó*?

Alkorpusz: sajtó, szépirodalom, tudományos, hivatalos, személyes
Lekérdezés: [word = "vörös"] ;
Találatok száma: 6135 db 39,89 db / millió szó
Keresési idő: 0,02s

1.	az ukrán lapok, végleg kihúzta a talajt a "	**vörös**	bárónak " nevezett kolhozelnökök lába alól. A megnevezés arra
2.	Szakértők megítélése szerint a NASA új Mars-programja vízválasztó lehet a	**vörös**	bolygó kutatásának eddigi történetében. Az új stratégia lehetőséget nyújt
3.	homályos ebédlőben sem oszlott Felicián feszült várakozása. Az üvegkancsó	**vörös**	bora hiába ragyogott reá. Barátom oly bizalmatlan volt,
4.	minden második francia húsétel receptje előírja, hogy a marhahúst	**vörös**	borban kell főzni, esetleg a bort a majdnem kész
5.	is éli meg nagyon élesen. Egyébként pedig amióta nincs	**vörös**	csillag a parlament kupoláján, nincs miért elmenni. Ami
6.	A tavaly augusztus 20-i tűzijátékon a látványelemek között több ötágú	**vörös**	csillag is látható volt kék körben. Emiatt két magyar
7.	láger. Gyökeresen más volt a helyzet, mikor a	**vörös**	hadsereg számára kerestek gépkocsikhoz értő szerelőket, lakatosokat. Oda
8.	egyben, Goldmann úr serénykedett, kövér, elefánttermetű,	**vörös**	hajú férfi. Kizárólag ő foglalkozott a betérőkkel. Nevéről
9.	Samphan és Nuon Chea felelősségre vonását. " A két	**vörös**	khmer vezető most Phnompenben tartózkodik. Miután mintegy kétmillió ember
10.	kapcsán érdemes felmelegíteni egy kicsit. Bár a topic a	**vörös**	maffiáról szól, jelen esetben a dolog nem ilyen egyszerű
11.	fogadja a Bodnár Lászlóval felálló Dinamo Kijevet. A "	**vörös**	ördögök " helyzetét nehezíti, hogy Michael Owen megsérült a
12.	vb döntőjébe jutás felé. Általános vélemény szerint a "	**vörös**	ördögök " nemcsak a sírból hozták vissza az egy pontot
13.	fogadj el tőlem jelképesen, az éteren keresztül egy szál	**vörös**	rózsát. Zon. Nem igazán értem a lányokat!
14.	a zöldmezős beruházás tönkreteszi a lankás vidéket, a takaros	**vörös**	téglás házsorok hangulatát. El kell ismerni, hogy a
15.	hőség volt, a torony hűvös és nyirkos, padlója	**vörös**	tégla, üvegtelen lőrésablakán esténként bejöttek a denevérek. Úgyhogy
16.	került sor. Az új párttitkár visszatette az egyetemre a	**vörös**	zászlót és a vörös csillagot. Mit ad Isten,

Találatok alkorpuszok szerinti megoszlása:

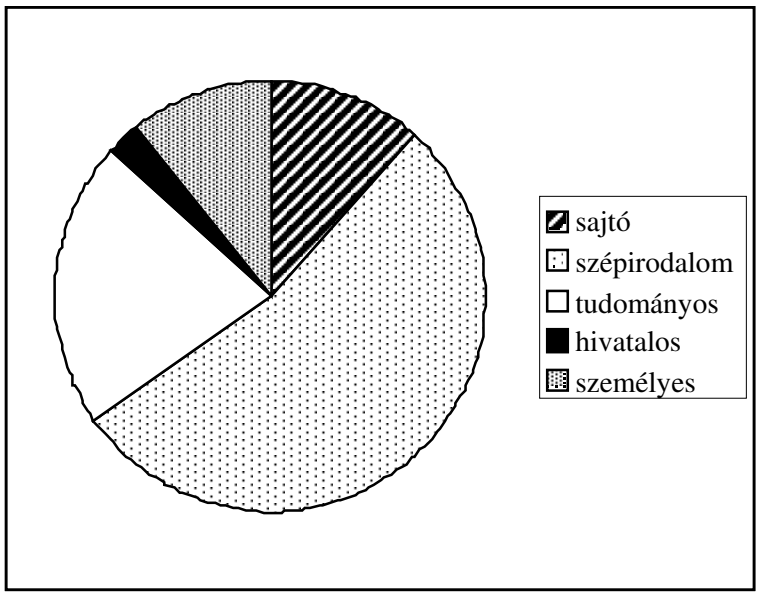
sajtó
2402 db
29,65 db / millió szó

szépirodalom
1997 db
137,29 db / millió szó

tudományos
1118 db
54,44 db / millió szó

hivatalos
142 db
7,15 db / millió szó

személyes
476 db
26,68 db / millió szó



65. ábra: A *vörös* keresési eredménye konkordanciapéldákkal az MNSZ-ben

Alkorpusz: sajtó, szépirodalom, tudományos, hivatalos, személyes
Lekérdezés: [word = "piros"] ;
Találatok száma: 8632 db 56,13 db / millió szó

1.	2) fekete 7-es busz, 3-as metró 3)	**piros**	7-es busz, 3-as metró 4) 47-es vill.,
2.	A hologram az interferenciákkal teli fénytérnek egy metszete, a	**piros**	a kis, a kék a nagy intenzitást jelenti.
3.	almám egyszerre pirosodott be. De hát tudjuk, a	**piros**	alma még nem érett alma, ne kezdjük el túl
4.	A kollégiumból kiözönlő " diákurak " néha bizony elcsentek egy	**piros**	almát vagy mézeskalács - szívet, mikor csínytevésből, mikor
5.	neuronjaimban, miért ver erősebben a szívem, miért lesz	**piros**	az arcom nem tudok információt szolgáltatni. * Nehézkes módszer
6.	Most már szinte naponta látom. Tolja maga előtt a	**piros**	babakocsit, aminek még a sátorteteje is fel van húzva
7.	állomás környékén dobálták el azok a felszabadult emberek az utolsó	**piros**	bankókat, akik a határ túlsó oldalán más pénzegységre vágytak
8.	, aztán bementek az egyik házba, s csakhamar egy	**piros**	bársonyfotellal tértek vissza. A fotelt a teherautó végéhez állították
9.	, hogy ezután is fognak azért együtt dolgozni. A	**piros**	bársonnyal bevont karosszéket - egész garnitúra volt belőlük - jól
10.	hétköznapi. Merthogy szürke mindennapokból jóval több van, mint	**piros**	betűs dátumokból. Eddig, s ne tovább! -
11.	annak idején 13 nappal. Az új Oroszországban ez már	**piros**	betűs ünnep, igaz csak egy napos. Az oroszok
12.	hogy gépkocsit lopjon. Balatonfüreden megtetszett nekik egy német rendszámú	**piros**	BMW és az elkötött járművel először egy gyümölcsösben rejtőztek el
13.	már a nap. Valakivel kószál a pára, aztán	**piros**	bogyók közt eltünik. Körünkből október varázsa csalja a sistergő
14.	Formába töltjük és hagyjuk megdermedni. A formából kiborítjuk és	**piros**	bogyós gyümölcsökből készült mártással kínáljuk. A vacsorához elengedhetetlena
15.	öreg miniszteri tanácsos másnap megállította a munkát. A térképen	**piros**	ceruzájával új részt jelölt meg a tervezett csatorna halvány kék
16.	a napon meg képes rögtön benáthásodni. Egy kivágott orrú	**piros**	cipőt legalább tíz percig nézünk, s végül is nem

Találatok alkorpuszok szerinti megoszlása:

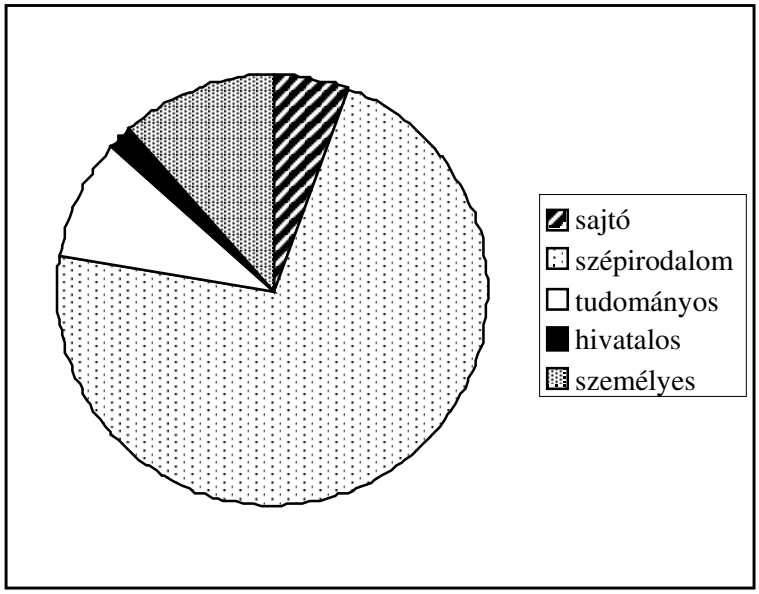
sajtó
2092 db
25,82 db / millió szó

szépirodalom
4642 db
319,12 db / millió szó

tudományos
834 db
40,61 db / millió szó

hivatalos
129 db
6,50 db / millió szó

személyes
935 db
52,41 db / millió szó



66. ábra: A *piros* eredménye konkordanciapéldákkal az MNSZ-ben

A feladatot tovább bővíthetjük és megkérdezhetjük a diákoktól, hogy a *vörö*sön és *piroson* kívül hallottak-e más kifejezést is, amit ezekre a színekre használnak? Előfordulhat, hogy nemleges a válasz, de lehet, hogy a *veres* szót is megemlíti valaki. Ezek után a *veres* konkordanciáit is össze lehet hasonlítani a *vörös*sel. A konkordanciák alapján a *veres bor, veres bársony, veres képű, veres tinta* kifejezések tűnnek ki leginkább, bár ma ez furcsának tűnhet.

Természetesen a kollokációk keresése és megfigyelése a magyart vagy más nyelvet idegen nyelvként tanuló diákok esetében is igen hasznos, hiszen még az idegen nyelvet magas szinten beszélők esetében is itt fordul elő a legtöbb hiba. Howarth (1996) lengyel anyanyelvű angoltanárok írásait vizsgálva hívta fel erre a figyelmet. A konkordanciák kiválasztásakor azonban a nyelvi szintnek megfelelők kiválasztása igen fontos. Célszerű a keresést csak egy bizonyos korpuszra korlátozni, például a sajtóra, ha a diákok célja a magyar nyelvű újságok megértése.

A konkordanciák nyomtatásakor igen könnyen ki lehet törölni a keresett szót, így az eredmény így néz majd ki:

bizonyította is, két vébé bronzérmet	Képességeik alapján a női párbajtörőzőinket
1931-ben teológiai doktorátust	Debrecenben. 1936-ban vallásoktatási igazgató
egy év alatt 1 millió dollárt	). Az IAAF elnökének örökségét felmérő cikkek
élvezet kiiktatása helyett élvezetet	. Az ő tablettáik legfeljebb a közérzetet javították. "
illusztrációival vált ismertté. Érdemeket	a rajztanítás megreformálása terén. Művei a
súlycsoportjában Csák József ezüstérmet	. Csák a pénteki három győzelme után vasárnap

a német légiós két gólt	, rajta kívül Kancselszkisz és van Bronckhorst
más fölött hatalmat	. Azokat, akik a pártszékházat védték, ha tényleg
hitelkártya-információkat	és pornográf anyagokat küldözgetett. -
Honnan	a pénzt az " előadásra ", hisz a sok
Innen	tudomást a legfelsőbb katonai vezetés is. A lap már
jogellenesen	javaik legalizálását szolgáló vállalkozásaik
polgári peres úton próbál jogorvoslatot	. HVG 98/38/98. szám 1998._szeptember_26.
valamennyi kategóriagyőztes jogot	a szeptemberi országos döntőn való indulásra.
erre a gyűjtésre is meg kellett	a pártközpont engedélyét, megmutattam a

67. ábra: Konkordanciák a keresett szó nélkül

A feladat itt az, hogy a hiányzó szót a szövegkörnyezet alapján találják ki a tanulók. Ebben olyan szókapcsolatok segíthetnek majd, mint a valamilyen *végzettséget* vagy *érmet szerez.* Tehát a kollokációk felismerését és tanulását segíti. Játékos gyorsasági versenyt is rendezhetünk ebből, több fordulóval. Ezek lehetnek egyéni vagy csapatversenyek, melyek könnyebb vagy nehezebb feladatokból állhatnak. A 67. ábra hiányzó szava a *szerez*, ami a mondatokban természetesen különböző alakokban szerepel. A példából azt is láthatjuk, hogy sokszor elég szinte csak egy szóra nézni, és azonnal tudjuk a választ: *érmet* és *gólt szerez*. Természetesen gólt *rúgni* is és *dobni* is lehet, és *érmet* is lehet *nyerni,* de a *szerez* az, amelyik a két szót együtt látva azonnal az eszünkbe jut.

Nehezebb kifejezések vagy idegen nyelv esetén sokat segíthet, ha egy listát készítünk, amelyből a diákok kiválaszthatják a helyes megoldást. Ez állhat két vagy több választási lehetőségből is. E példa esetében a *szerez* és a *nyer* megfelelőnek tűnnek választási lehetőségként.

Az ezen az elven készített feladatok sokaságát lehet kitalálni. Például két vagy több keresés eredményét „összekeverve” a kulcsszó helyét üresen hagyva, a diákoknak kell megállapítani, hogy melyik mondatba melyik szó illik, vagy a szót milyen formában kell használni (igeragozás, igeidők, egyes-többes szám stb).

A nyelvtanulás egy másik nehézségéről, a homonímiáról is szólnunk kell. A magyarban az *ég, égő* vagy *vár* szerepel tipikusan a homonimák példájaként a tankönyvben. Más nyelvekben is nehézségeket okoznak az ilyen szavak, az angolban meglehetősen sok is van belőlük. A homonimák jelentésének felismerésében nagy szerepet játszik a szövegkörnyezet és a kollokációk ismerete. Nézzük meg az angol *flat* szó konkordanciáit, és készítsünk listát különböző jelentéseiről a leggyakrabban velük együtt előforduló szavakkal együtt. A kereséskor a keresett szó előtti és utáni szavak ábécé sorrendben szerepelnek. A legtöbb esetben e három szó elegendő a jelentés felismeréséhez.

```
◄n her own for the previous year in a flat above a small craft shop that she ra┼
◄roughton's production, which makes a flat debut for the Birmingham Rep Company┼
◄rson pressed hard for a `premium", a flat deduction of 1s. a week from all une┼
◄ndi, are not isolated mountains in a flat desert landscape. They have an influ┼
◄ment, the banning of `he that hath a flat nose", and the writing `the fathers ┼
◄hares finished a traumatic week on a flat note with the FT-SE index, measuring┼
◄ blamed for the setback. <p> After a flat opening beers perked up when it beca┼
◄d boorish Vic, face in repose like a flat tyre. Haydn Gwynne, looking uncannil┼
◄ expectations. Sales remained almost flat at Ł626m. The final dividend is unch┼
```

◄editated art, stressing contours **and flat tones**. Friends of Gauguin had organi┼
◄mediately present one of colours **and flat two-dimensional** shapes. When we say ┼
◄hen the regions behind the waves **are flat. These** flat regions must be describe┼
◄rsts of crossfire gabble in **Ashley's flat-voiced singspeak**. Even when an audib┼
◄erkin's performances of Schubert's **B flat major** sonata, Wanderer fantasy and t┼
◄qually outstanding Piano Sonata in **B flat minor** has yet to find favour with pi┼
◄huffing up the stairs of her **council flat,** berating her indolent husband for b┼
◄ a night-time raid on the **deceased's flat,** during which they attacked and kill┼
◄k his BMW beside the flash **docklands flat,** close to his rough but pure roots, ┼
◄ng superbly to the final altissimo **E flat _ a** stratospheric note which she the┼
◄d Liszt's Piano Concierto No. 1 in **E flat as** `The triangle concierto" because ┼
◄s year at Esterhaza, the Sonata in **E Flat (Hob** XVI:49). The difference is stri┼
◄ it's better to try _" <p> `And **fall flat on** your face?" <p> `You're jumping t┼
◄pirouettes and arabesques _ I'd **fall flat on** my face" but equally doesn't want┼
◄> <hl> Wace Group rights issue **falls flat </hl>** <bl> By ALISON EADIE </bl> <st┼
◄d Peter Walsh, who rode in the **first Flat race** trial there yesterday says: `I ┼
◄ack at Southwell is put on trial **for Flat and** National Hunt stables and the su┼
◄the amiable Richard Muddle, **a former Flat jockey**, who allied his vast internat┼
◄of a social occasion at **Mrs Gorman's flat,** attributing to her an `unseemly and┼
◄A stocking footprint in blood at **her flat was** consistent with the size and sha┼
◄pondent opted to live in a **high-rise flat,** most often confused with Sixties pr┼
◄nd greater poverty." <p> He said **his flat on** the estate had been robbed by a m┼
◄005 </dt> <hl> Deaf woman stabbed **in flat </hl>** <bl> By JACK O'SULLIVAN </bl> ┼
◄ Steve Cauthen may dominate the **last Flat meeting** of the season at the Berkshi┼
◄lthy, appetizing meal in ten **minutes flat,** you should be able to do so. Health┼
◄e with the Pattern, the programme **of Flat racing** adopted in 1965 to test the b┼
◄a marketing problem in offering **only Flat racing** during the summer months, whi┼
◄ to begin with a simple, relatively **`flat" problem** (in our case it was a large┼
◄nd Jockeys Making The Running on **The Flat </hl>** <st> <sect> Sport Page 29 </se┼
◄ircus ring unpainted, permitting **the flat tone** of his white ground to serve as┼
◄consistencies in character or **theme, flat and** functional dialogue, awkward sel┼
ther was hot and the pitch **typically flat."** <p> The only disappointment for th┼
ough the sun shone and the water **was flat,** a slight head wind blunted the time┼
 store-for-store basis as volume **was flat and** prices fell. <p> Michael Pickard┼
eneral Instruments, which is **working flat out** on it at its plant in San Diego, ┼

68. ábra: A *flat* konkordanciája

A következő feladatban a cél az, hogy a hiányzó szót az azt követő példák alapján találja ki a diák. Természetesen itt azt várjuk, hogy a diák ismerje fel az antibiotikum, heroin és aszpirin szót látva, hogy ezek az angol *drugs* jelentéskörébe tartoznak (magyarul viszont nem tudjuk a feladatot egy szóval megoldani), így a hiányzó szó a *drugs*.

rough less over-prescribing of _____ such as antibiotics and laxatives, and pounds 45 m
    in a pure state. Research into _____ such as heroin and cocaine has suffered for simila
    It was known that antiplatelet _____ such as aspirin could sometimes help, but doctors

69. ábra: Feladat a szuperordinátok és hiponímák gyakorlására

Szemantikai prozódia[32]

Sokan talán még nem is hallották ezt a kifejezést, de azonnal rájönnek, hogy mit jelent, ha csak arra a két kifejezésre gondolnak, hogy *örömet ???* és *bánatot ???*. Ugye nem okozott gondot ennek a megfejtése? Az ehhez hasonló kifejezéseket gondolkodás nélkül használjuk az anyanyelvünkön, de amíg nem vizsgáljuk meg tudatosan a szövegkörnyezetet, nem tudjuk megmondani, hogy miért pont a *szerez* vagy az *okoz* kifejezéseket használjuk egyik vagy a másik esetben. Idegen nyelv esetén, a helyes nyelvhasználatot csak tanulással érhetjük el. A magyar nyelvben talán érezzük, hogy az *örömet szerezni* szoktuk, a *bánatot* pedig *okozni*. Az MNSZ-ben mind az *örömöt*, mind pedig az *örömet* szót megkerestem, és a tőle jobbra álló szavakat ábécérendbe raktam. Az első érdekes megfigyelés az volt, hogy az *örömet* 1086, az *örömöt* pedig 694 alkalommal fordult elő. Az *örömet* listát megfigyelve előfordult ugyan az *okoz* ige is, de egyrészt itt több esetben tagadással együtt fordult elő: *másoknak ez* ***nem örömet*** *okoz*, *hanem szenvedést,* másrészt a *szerez* kifejezéssel sokkal több konkordanciát találtunk. Mivel nem lehet az összes konkordanciát lekérni, így a hamis kép elkerülése érdekében, és a hipotézis igazolására a *bánat* és *szerez* szótövekre keresve kaptunk ugyan három konkordanciát, de alaposabban megnézve rá kell jönnünk, hogy az első mondatban a *szerez* igazából az *örömre* vonatkozik. A második esetben az *orosz bánattal* végződik a mondat, és csak a következő mondatban szerepel a *szerez*. A harmadik példában *szerez felszabadultságot,* és nem *bánatot* az összetartozó rész.

Alkorpusz: sajtó, szépirodalom, tudományos, hivatalos, személyes
Lekérdezés: ([lemma = "bánat"] []{0,5} [lemma = "szerez"]) | ([lemma = "szerez"] []{0,5} [lemma = "bánat"]) ;
Találatok száma: 3 db 0,02 db / millió szó

1.	nem tudhatja, mikor mi	**szerez** *V.e3*	majd örömet vagy **bánatot** *N.ACC*. Számomra az
2.	- Amerikai öröm, orosz	**bánat** *N.NOM*	Meglepetésre Michelle Kwan **szerezte** *V.TMe3* meg a
3.	és tapossa ki magából a	**bánatot** *N.ACC*	, a keserűséget, **szerez** *V.e3* olcsó felszabadultságot.

70. ábra: A *bánat* és *szerez* lekérdezése az MNSZ-ből

Alkorpusz: sajtó, szépirodalom, tudományos, hivatalos, személyes
Lekérdezés: ([lemma = "bánat"] []{0,5} [lemma = "okoz"]) | ([lemma = "okoz"] []{0,5} [lemma = "bánat"]) ;
Találatok száma: 38 db 0,25 db / millió szó

1.	szerető atyámtól. Hogy én	**okoztam** *V.Me1*	a mahárádzsának a **bánatát** *N.PSe3.ACC*: bánt, hogy
2.	a melegből. Talán nem	**okozok** *V.e1*	**bánatot** *N.ACC*, ha mégis visszagondolok első
3.	? - Mindenkinek bajt meg	**bánatot** *N.ACC*	kell **okozni** *V.INF*. Az nagyon rossz lehet
4.	mulat, de hogy sok	**bánatot** *N.ACC*	nem **okozott** *V.Me3* neki rendőrség, kormányellenőrzés,
5.	múlt el nap, hogy	**bánatot** *N.ACC*	ne **okozott** *V.Me3* volna valakinek a szeleburdiságával.
6.	múlt el nap, hogy	**bánatot** *N.ACC*	ne **okozott** *V.Me3* volna valakinek a szeleburdiságával.
7.	egy dolgot, nem akarok	**bánatot** *N.ACC*	**okozni** *V.INF* magának, de hány éves
8.	mondta: - Nem akartam	**bánatot** *N.ACC*	**okozni** *V.INF* senkinek. Várakozva, keményen
9.	belül örök örömöket és örök	**bánatokat** *N.PL.ACC*	**okoz** *V.e3* az örök szerelem. Az

[32] E kifejezéssel Louw (1993) írásában találkozhatunk először, de egyre több szerző használja művében, így tehát szakkifejezésként egyre elfogadottabbá válik.

10. (4) Keserűséget, **bánatot** *N.ACC* **okoz** *V.e3* vkinek. Keserítette sok bú
11. , hogy ő most ekkora **bánatot** *N.ACC* **okozott** *V.Me3* a kisfiúnak is meg a
12. , hogy ő most ekkora **ánatot** *N.ACC* **okozott** *V.Me3* a kisfiúnak is meg a
13. zsidó volt, aki sok **bánatot** *N.ACC* **okozott** *V.Me3* neki túlzott csillogni akarásával,
14. az egykori állampárt bajt és **bánatot** *N.ACC* **okozott** *V.Me3*, és annyi tanulságot azért
15. Nagyon fájt, hogy **bánatot** *N.ACC* **okoztam** *V.TMe1* neki, el is döntöttem
16. A társaságok mérlegadatai nem **okoztak** *V.Mt3* sem váratlan örömöt, sem **bánatot** *N.ACC* az
17. érzéseivel, de Iunior már **bánatot** *N.ACC* se tudott neki **okozni** *V.INF*, valahogy közömbös lett.
18. Vörös Piroska nevezetű volt kedvese **okozta** *V.TMe3* szerelmi **bánatát** *N.PSe3.ACC* úgy
19. is mindaz ellen, ami **bánatot** *N.ACC* gyötrődést **okoz** *V.e3* a nagyvilágban és a kisvilágban
20. , másutt pedig a felhők **bánatot** *N.ACC* , kiábrándulást **okoztak** *V.Mt3*. Sokunkat nevetségessé

71. ábra: A *bánat* és *okoz* lekérdezése az MNSZ-ből

A fenti konkordanciákban is találunk további példákat arra, hogy *kiábrándulást, gyötrődést okoz* valaki vagy valami. Tehát vannak olyan igék, amelyeket csak bizonyos negatív jelentésű főnevekkel, másokat meg pozitívakkal használunk. Érdekes megjegyezni azt is, hogy az *okoz* angol és német megfelelői is hasonlóképpen viselkednek. Az alábbi példa a *verursachen* szemantikai prozódiáját vizsgálja.

Forrás: Bonner Zeitungskorpus (IDS Mannheim)
Szoftver: Tim Johns & Mike Scott, *Microconcord*
A találatokat a kulcsszót megelőző szavak ábécérendjébe állítva láthatjuk
Vizsgálat tárgya: szemantikai prozódia

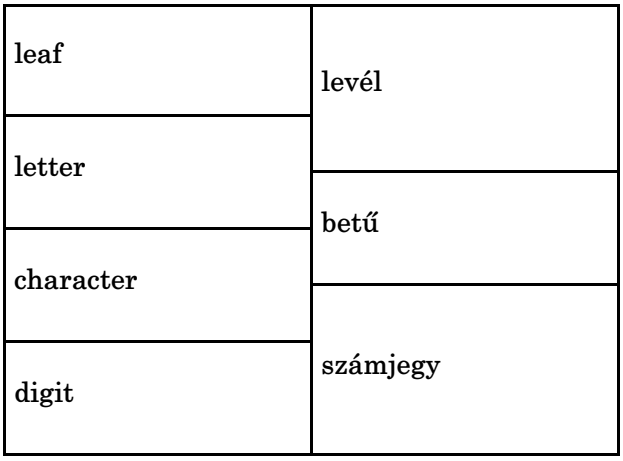
1 muß, dann wehe denen, die das Ärgernis verursachen. (Acheson spielte hiermit auf ei
2 ld von Weber registrierten Abweichungen verursacht sein. Michael Globig. Mehr
3 -Grenzkommandos beträchtliche Aufregung verursacht. Bei Redaktionsschluß war
4 Prozent einen Verlust von 5504 DM aus, verursacht durch erhöhte Steuerzahlungen.
5 um Teil durch die schärfere Besteuerung verursacht. Die HV soll am 6. Mai eine auf
6 äden und Verluste unter der Bevölkerung verursacht haben. (Reuter - dpd - AP - UP).
7 sie sonst noch darunter? Messer? Damit verursachte man weniger Lärm als mit einem
8 gab der Münchener Rundfunk bekannt. Der verursachte Sachschaden beträgt 4,8
9 rnommenen Rentenposten von 135 Mill. DM verursacht wurden. Dieser riesige Verlust
10 en kann. den durch die Drei-Tage-Woche verursachten volkswirtschaftlichen Verluste

72. ábra: A *verursachen* szemantikai prozódiája (Bill Dodd honlapjáról)

Fordítási ekvivalenciák

A rossz beidegződések sokáig megmaradnak, és sok diák fejében él az egy magyar szó = egy idegen szó képzet. E felfogás módosítására lehet felhasználni a következő konkordanciákra épülő feladatot. Az angol *leaf* szó konkordanciái közül válasszuk ki azokat, amelyek a (fa)levél jelentésben szerepelnek. A konkordanciákban szereplő kulcsszót fordítsák le a diákok. Ezek után nyomtassuk ki a *levél* konkordanciáit, ügyelve arra, hogy ne csak az előbbi értelemben szerepeljen. A fordításokban szerepelni fog a *letter.* A *letter* konkordanciáit úgy válogassuk ki, hogy legyen benne *betű*vel fordítható

is. Folytathatnánk a sort a végtelenségig. A fordítások eredményét foglalják a diákok valamilyen rendszerbe. A 73. ábra egy ilyen lehetséges ábrázolást mutat be.



73. ábra: A fordítási ekvivalensek egyik ábrázolási módja

5.5.2. A „számok tükrében”

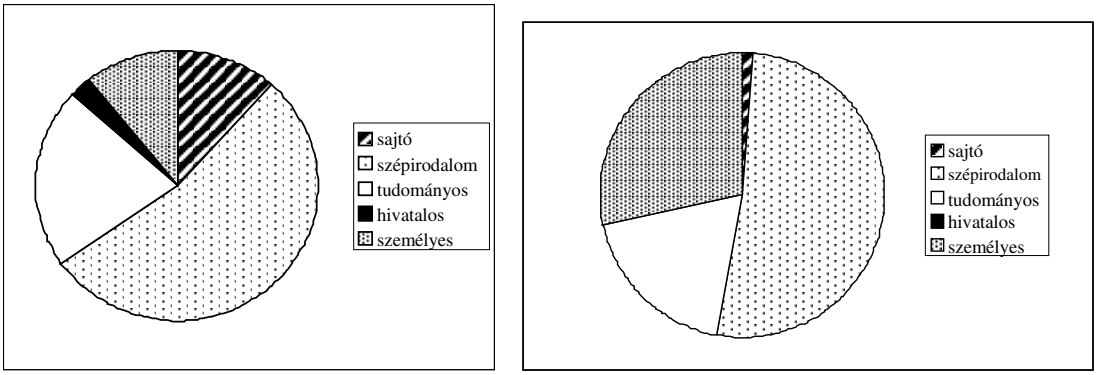
A konkordanciák alatt, mind a 65. ábra és a 66. ábra, az alkorpuszok szerinti megoszlás gyakoriságát és arányát is szemlélteti. A *vörös* 6135 alkalommal fordult elő a teljes korpuszban, ami 39,89 alkalmat jelent egymillió szavanként. Ugyanebben a korpuszban a *veres* 396 alkalommal fordul elő és ez 2,58 db/millió szót jelent. Mivel ugyanazt a korpuszt használtuk mindkét keresés esetén, elég a puszta számokat összehasonlítani, hogy megállapíthassuk: a *veres* változatot sokkal kevesebbet használjuk, mint a *vörös*et. Az alkorpuszok vizsgálata alkalmat ad arra, hogy a diákok a különböző zsánerek stílusjegyeit, szókincsét alaposabban megvizsgálhassák. Tehát a következő kérdéseket lehet feltenni:

- Melyik alkorpuszban szerepel a *vörös* és *a veres* leggyakrabban, és mi lehet az oka?
- Mire következtetünk abból, hogy a hivatalos szövegekben egyszer sem, és a sajtóban is nagyon kis számban fordul elő a *veres*?

Ha az alkorpuszokbeli megoszlást nézzük, akkor láthatjuk, hogy a *vörös* legtöbbször a sajtóban fordul elő, aminek nyilvánvaló oka, hogy politikai értelemben használják. A *vörös* a megszokott köznyelvi változat, így nem meglepő, hogy a hivatalos szövegekben egyszer sem, és a sajtóban is elenyésző számban szerepel.

VÖRÖS		
sajtó	2402 db	29,65 db / millió szó
szépirodalom	1997 db	137,29 db / millió szó
tudományos	1118 db	54,44 db / millió szó
hivatalos	142 db	7,15 db / millió szó
személyes	476 db	26,68 db / millió szó

VERES		
sajtó	22 db	0,27 db / millió szó
szépirodalom	171 db	11,76 db / millió szó
tudományos	88 db	4,29 db / millió szó
hivatalos	0 db	0 db / millió szó
személyes	115 db	6,45 db / millió szó



74. ábra: A *vörös* és a *veres* gyakoriságának összehasonlítása

Itt kell azt is megjegyeznünk, ha a vizsgálatokhoz használt korpuszok mérete különbözik, akkor teljesen hamis képet adna, ha csak az előfordulások számát hasonlítanánk össze. Így a *vörös* előfordulását megfigyelve láthatjuk, hogy a sajtóban 2402 alkalommal, a szépirodalomban pedig csak 1997 alkalommal fordul elő. Ha azonban az egy millió szóra eső előfordulást nézzük, azonnal látjuk, hogy a szépirodalomban négyszer olyan gyakran fordul elő, mint a sajtóban. Tehát maga a szépirodalmi korpusz jóval kisebb, mint a sajtó korpusza. Ha valaki saját maga készít korpuszt, akkor valószínű, hogy kisebb egységekre vetítve kell normalizálnia a kapott eredményeket, például 100 000 szóra vetítve. A kördiagram az MNSZ estében ezeket a normalizált arányokat mutatja.

A számok vizsgálata, tehát a statisztikai adatok vizsgálata az anyanyelvi beszélők esetében is fontos, mivel a nyelvi intuíció alapján az anyanyelvi beszélő meg tudja ugyan mondani, hogy egy mondat vagy szerkezet helyes-e, de a gyakoriság esetében az intuíció nem segít. A találgatásban a nem anyanyelvi beszélő is ugyanolyan esélyekkel indul tehát.

A különböző zsánerek, regiszterek vizsgálatakor megfigyelhetjük egy bizonyos szó vagy szerkezet előfordulásának gyakoriságát, vagy a leggyakrabban előforduló szavak listáját is összehasonlíthatjuk. Így például a beszélt és írott nyelv korpuszában leggyakrabban előforduló szavakat, vagy a szenvedő szerkezetek számát vizsgálva bizonyos következtetéseket vonhatunk le. Ezek egyben mintaként is szolgálnak majd hasonló stílusú szövegek alkotásához.

5.6. Számítógépes feladatok

Az eddig említett feladatok mindegyikét maguk a diákok is elvégezhetik a számítógép használatával. A számítógép használatának megkezdésekor érdemes olyan feladatokkal kezdeni, amelyeket nyomtatott változatból már ismernek a diákok. A számítógép használatának az az előnye, hogy a diákok saját maguk próbálhatják ki, hogy például hogyan változnak a vizuálisan könnyen felismerhető minták, ha a keresett szótól jobbra vagy balra eső szó alapján rendezzük a konkordanciákat ábécérendbe. Ha valamire felfigyel-

nek és ellenőrizni akarják, akkor azt azonnal megtehetik, és nem kell arra várni, hogy a tanár a következő órára kinyomtassa.

Vannak azonban olyan programok, amelyek csakis ilyen interaktív környezetben képzelhetők el, vagy hatásosak. Itt sokakban felmerülhet a kétség, hogy kisgyerekekkel nem lehet számítógépen az itt leírtakhoz hasonló feladatokat végezni. Paul Thompson, Alison Sealey és Mike Scott a TaLC 6. konferenciáján tartott előadást (2004) egy folyamatban levő kísérletről, amelyben két alsó tagozatos iskolai osztály (8-9 és 9-10 éves gyerekek) tanulói vesznek részt. A kísérlet során ők is a konkordanciák nyomtatott változatával kezdték, és a Wordsmith Tools program módosított kezelői felületét használták a számítógépes feladatokhoz a későbbiekben. Elsősorban a szófajok disztribúcióit és szinonimákat vizsgáltak a gyerekek által feltett kérdések mellett.

Contexts

A Contexts (Johns, 1994) olyan számítógépes program, amely konkordanciákra épül. A program létrejöttéről a *Teaching and Language Corpora* című könyvben olvashatunk bővebben (Johns, 1997). A programot lehet megfigyelésre és tanulásra vagy kvíz formában való játékos tesztelésre használni. A kvízszerű használat tűnik a legérdekesebbnek, és a diákok önállóan is használhatják. A program az angol nyelv kollokációinak elsajátítását szándékozik elősegíteni, így menüi angol nyelvűek. Egy több mint 1000 kulcsszóra készült feladatsor is tartozik hozzá, amelyet a programmal együtt automatikusan letöltünk. A program szerkeszthető, ami azt jelenti, hogy ha más nyelv tanulására kívánjuk felhasználni, akkor a benne szereplő adatokat az általunk választott nyelvre kicserélhetjük. Ez persze jelentős időbefektetést igényel.

A CONTEXTS-ben szereplő kulcsszavak mindegyikéhez 10–10 konkordancia tartozik, melyeket az angol esetében már készen kapunk, de más nyelvek esetében magunknak kell kiválasztani és megszerkeszteni. A kvíz indításakor csak egy konkordanciát látunk, melyből hiányzik a kulcsszó. Egy konkordancia esetében talán több szó is beillik a szövegbe, így legtöbbször nem elég egy konkordancia megtekintése a helyes megoldás kitalálásához. Minden további konkordanciát egy gomb megnyomásával kell „kérni". Bármikor lehet tippelni a hiányzó szóra, és ha nem sikerült eltalálnunk, akkor folytathatjuk a következő konkordancia megtekintésével. Bármikor kérhetünk egy kis segítséget, ami azt jelenti, hogy az egy kvízbe tartozó 10 kulcsszó mindegyike megjelenik a képernyőn, így csak azon kell gondolkodni, hogy ezek közül melyik illik a szövegbe.

A legtöbb számítógépes programmal az a baj, hogy már a legkisebb eltérés esetén is, például ha a kisbetű helyett nagybetűt használunk, hibásnak tekintheti a megoldást. Ez a program azonban a helyesírási hibákra is „intelligensen" reagál. A kvíz végén összefoglalja és értékeli teljesítményünket, majd tanácsot ad a további tanulást illetően. Természetesen ez is angol nyelven olvasható.

A programban szereplő konkordanciák megváltoztatásához a következő lépések szükségesek:

Válasszuk ki a megfelelő kulcsszavakat, és minden kulcsszóhoz válasszunk ki 10 megfelelő konkordanciát. Ezeket másoljuk át a Wordpad programba a következő formát követve:

```
Body Parts
9000
Roberto Mussi has recovered from an ankle injury and may replace Mauro Tassotti
 driving rain, gale-force winds and ankle deep mud - conditions not unfamiliar a
sian philosopher of art, wore brown ankle boots with a zip. "I will not wear bla
e other women were all competing in ankle-length dresses and elaborate hats. But
ing my autograph book. I twisted my ankle as I landed, but I kept going, limping
blouses, very short pleated skirts, ankle socks and strappy sandals. Commentator
er who tied a rope around his son's ankles and dangled him upside down over the
er needing treatment for a sprained ankle. But now Mr Whittaker, 55, a former co
e throat, legs browned from knee to ankle - and all the rest startling white. Th
football, where people kick at your ankles. They share out resin for your hands
*
 168
ll. Young females walk along, often arm in arm, not going anywhere in particular
ding the country together by strong-arm methods. The results are a vindication of
ning forward in his chair, with his arms folded across his chest and staring at t
up in restful postures, folding our arms and crossing our legs; we sit, stand, sq
nt were holding the animal in their arms as if it were a baby. In 11 per cent of
ed round the car to take his wife's arm. "Come, my dear, we are going home," he t
set the idealists' beliefs that all arms could be laid aside at once, and while t
ary of Peking's overseas investment arm, the China International Trust and Invest
ntered white men. The raising of an arm in some form of Hail or Wave salute is al
d nurses tend to keep each other at arm's length; no wonder the Ren had trouble c
*
```

75. ábra: Konkordanciák szerkesztése a *Contexts* programhoz

Az első sorban a cím szerepel (Body Parts), az alatta levő sorba egy számot írunk vagy üresen marad. A kulcsszó a szövegfájl 37. oszlopában kezdődik és minden konkordancia 80 betűből áll, ha a kulcsszó nyolc betű vagy annál rövidebb. Minden kulcsszóhoz 10 konkordancia tartozik, és minden fájlban legalább 10 kulcsszó szerepel. Minden egységet egy új sorban levő * zár le, és az utolsó egység után *** áll.

Ha a konkordanciák mellé utasításokat vagy feladatokat akarunk írni, akkor a következőt kell tennünk: a konkordanciák után a * helyett egy @ jel közbeiktatásával kell ezeket megadnunk úgy, hogy ne legyen több hat sornál, és ezután kell a * jelet tenni.

```
@
According to these contexts, what sorts of things can you acquire?
Which of the things you can acquire are Concrete and which Abstract?
How would you try to acquire something "by association"?
In these contexts where could you not use "get" instead of "acquire"?
#I hope I shall never acquire _.
#It takes a long time to acquire _.
*
```

76. ábra: Utasítások szerkesztése a *Contexts* programhoz

Az így elkészített szövegfájlt .con kiterjesztésű fájlra kell változtatni ahhoz, hogy a program felismerje. Ezek után a Contexts.ini fájlt kell módosítani. A hatodik sorban a Make Index=FALSE sort kell TRUE-ra változtatni és a fájlt így elmenteni.

Network=TRUE
Log Quiz=FALSE
Quiz Extension=QUZ
Log Session=FALSE
Session Extension=SES
Make Index=FALSE
Log Path=C:\CONLOG
Line Printer=LPT1

77. ábra: A Contexts.ini fájl tartalma

A program futtatása után ezt az értéket változtassuk vissza az eredeti FALSE-ra.

A program eredetileg MS-DOS-ra készült, de a Windows 2000-es és XP változatán is jól működik. Maga a program tömörítve mindössze 529kb.

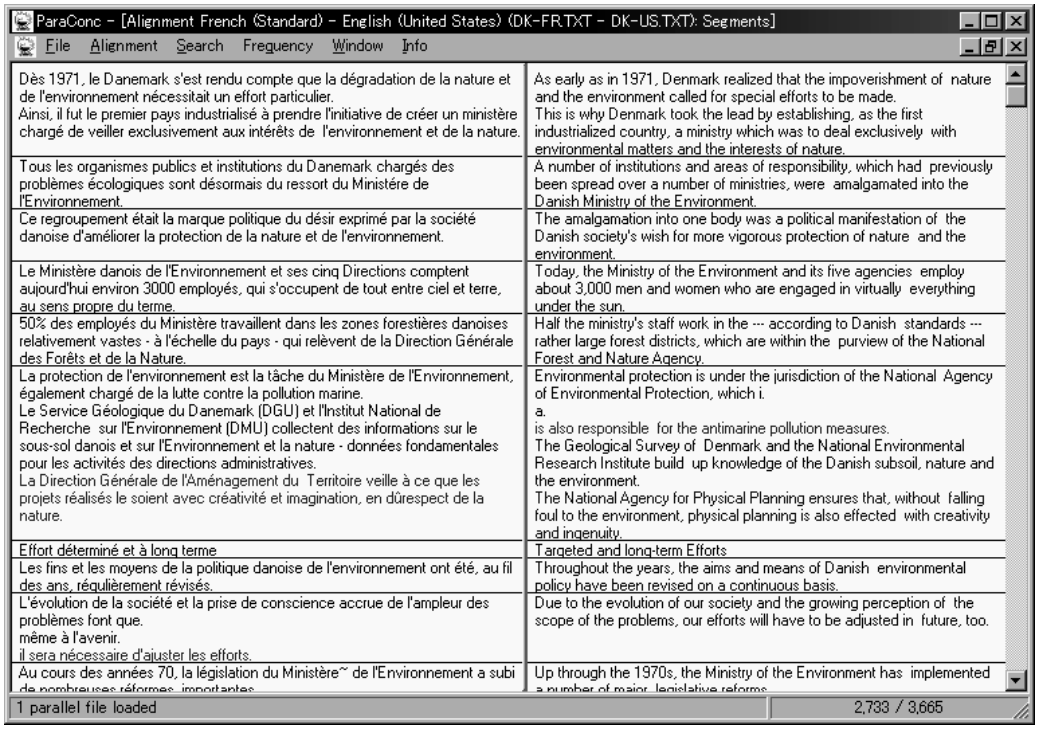
Interaktív címkéző program használata

A címkéző programok arra valók, hogy a szavak mellett feltüntessék a szófajt, és egyéb nyelvtani jellemzőket. Tony McEnery egy előadásában (AILA 2001, Tokyo) megemlítette, hogy az egyetemi nyelvoktatásban nagy gondot okoz a nyelvtan, mivel az angol közoktatásban vagy 20 éven keresztül nem tanították. Így még az angol szakos hallgatók sem ismerték azt megfelelően. Az önálló gyakorlás és önellenőrzés legegyszerűbb módja az volt, hogy egy taggert használtak fel erre, amely a téves megoldásokat visszajelezte. A diákok szívesen használták a nyelvtani hiányosságok bepótlásának ezt a módját.

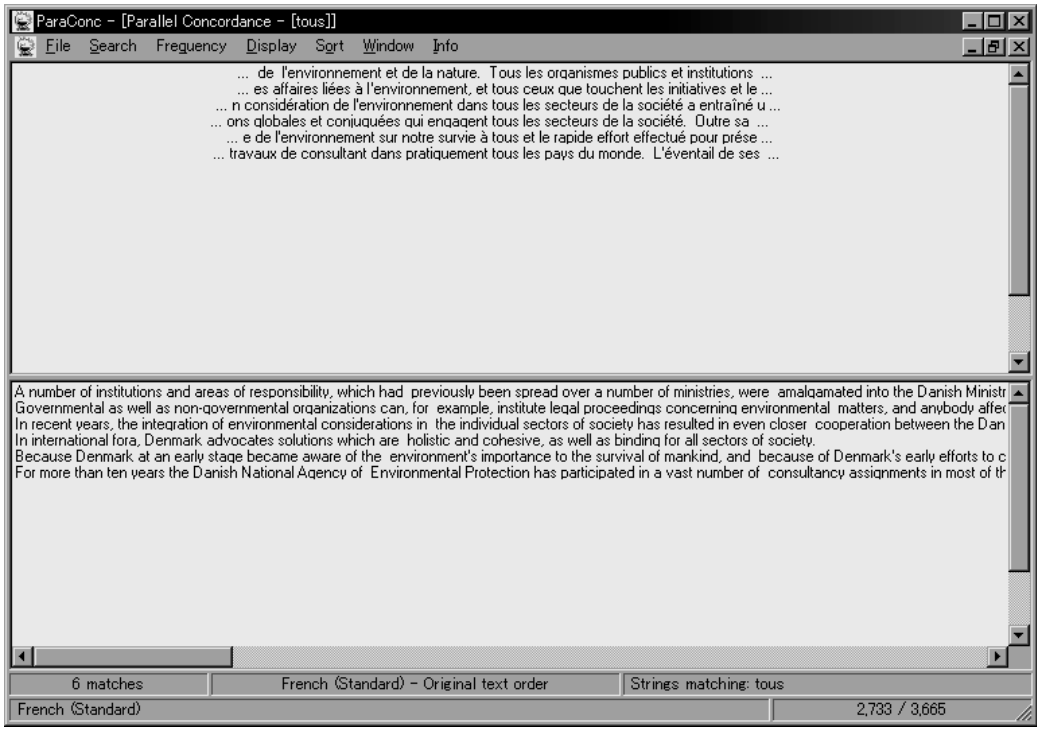
Párhuzamos konkordanciák

Mind a fordítók munkája, mind pedig a fordítóképzés elképzelhetetlen lenne számítógépes programok nélkül. Félreértés ne essék, nem a gépi fordítóprogramokra gondolunk, hanem a fordítást segítő programokra, mint például memóriabankok, amelyekben a már lefordított, és gyakran előforduló szakkifejezéseket, vagy terminus technicusokat tárolják a fordítók. Ezen programok mellett a párhuzamos konkordanciák is segítenek, elsősorban a fordítók képzésében. A 78. ábra az ilyen párhuzamos elrendezést szemlélteti (ParaConc program). A szövegekben tetszés szerint kereshetünk akár az angol, akár a francia szavak alapján. A keresett szót KWICK formátumban látjuk, a fordításokat pedig az ablak alsó részében teljes mondatok formájában (79. ábra). A két szöveg összehasonlítását támogató statisztikai funkciókat is elvégezhetünk, többek között disztribúciót vagy gyakoriságot vizsgálhatunk. A 80. ábra a gyakoriságra vonatkozó adatokat mutatja.

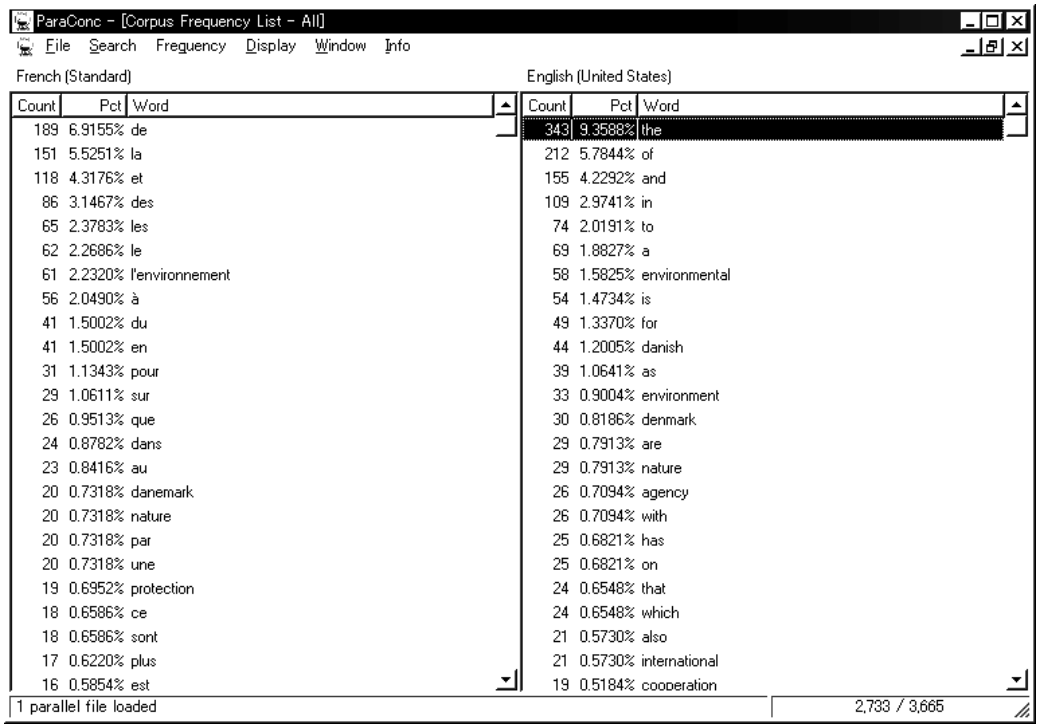
Attól függően, hogy milyen korpuszt használunk, vizsgálhatjuk az azonos művek különböző fordításait egy nyelven, vagy egyszerre több nyelven. A fordítás szempontjából az egyik legalaposabban elemzett mű a Biblia, melyhez viszonylag könnyen hozzá lehet férni különböző nyelveken elektronikus formában. Érdemes az ilyen vizsgálatokat klasszikus műveken végezni, hiszen azoknak sok nyelven létezik fordításuk, van úgy, hogy egy nyelven belül több is. A klasszikusok általában szabadon letölthetők az internetről, mert nem fűződnek hozzájuk copyright jogok.



78. ábra: Francia–angol párhuzamos szöveg



79. ábra: A *tous* konkordanciái és fordításai



80. ábra: Parallel szövegek gyakorisági listája

Végül azt is meg kell említenünk, hogy a korpusz hatalmas tárháza a kulturális jellegű információknak. Nem csak azért, mert olyan speciális kifejezések fordulhatnak elő például az angol különböző változataiban, amelyeket összevethetünk (pl. *tap, faucet*), hanem azért is, mert olyan eseményekre, művekre való utalások szerepelnek benne, amelyeket a tankönyvek nem említenek, de a köznapi beszédben gyakran használják.

5.7. Összefoglalás

Köztudott, hogy idő kell ahhoz, hogy a tudományos kutatás eredményei bekerüljenek az oktatásba. A számítógépek és az internet használatának széles körű elterjedését látva, és az európai közösségbe való tartozáshoz szükséges nyelvtanulás fontosságát figyelembe véve mindenképpen alkalmasnak tűnik az idő arra, hogy új lendületet kapjon a nyelvoktatás és nyelvtanulás hazánkban. A korpusznyelvészet szakirodalma szinte kizárólag angol nyelvű, így csak kevesen juthattak hozzá eddig az erre vonatkozó információkhoz.

Sajnos elkerülhetetlen, hogy a javasolt irodalom is angol nyelvű. A korpuszra épülő szótárak és segédanyagok bemutatásával példát szeretnénk mutatni arra, hogy a magyar nyelvet tanítók és kutatók milyen anyagok készítésével járulhatnának hozzá a magyart idegen nyelvként tanuló diákok eredményességéhez és az anyanyelv oktatásához.

A saját készítésű feladatok leírása szintén arra szolgált, hogy ötleteket adjon, még ha idegen nyelvű példákat is mutattunk be sok esetben. Láthattuk, hogy nem feltétlenül szükséges a számítógép használata a tanórákon, de igazából a végső cél mégis csak az,

hogy a diákok maguk kezelhessék a programokat és saját kis „kutatásaikat” végezzék. Ez különösen igaz a felsőoktatásra, ahol a jövő fordítóinak, pedagógusainak és nyelvvel foglalkozó szakembereinek feltétlenül meg kell ismerniük a korpuszok használatát és az oktatásban való felhasználás lehetőségeit és módszereit.

E könyvből minden bizonnyal sok információ kimaradt, a honlapok esetleg változtak, és új programok jelentek meg a kézirat lezárása után, ezért egy magyar nyelvű honlapot készítettünk, ahol a korpusz iránt érdeklődők magyar nyelven olvashatnak további információkat a www.korpusz.com honlapon. Kérdéseiket és megjegyzéseiket a korpusz@hotmail.com címen várjuk.

A KÖNYVBEN SZEREPLŐ NYELVI KORPUSZOK, SZÖVEGTÁRAK ÉS ADATBÁZISOK

(A korpuszok neveit legtöbbször eredeti formában adjuk meg. Amennyiben a korpusz honlappal rendelkezik, annak címét közöljük.)

American National Corpus; Amerikai Nemzeti Korpusz; http://americannationalcorpus.org/

American Printing House for the Blind (APHB)

Australian Corpus of English (ACE), amely Macquarie Corpus of Written Australian English néven is ismeretes. Kézikönyv: http://khnt.hit.uib.no/icame/manuals/ace/INDEX.HTM

Bank of English; http://www.titania.bham.ac.uk/docs/; A Collins Word Web része

Base de données textuelles ChroQué; http://www.tlfq.ulaval.ca/lexique/chroque/

Bergen Corpus of London Teenager Language (COLT); Londoni Tinédzser Nyelv Bergeni Korpusza; http://torvald.aksis.uib.no/colt/

BESEDA (szlovén nyelv); http://bos.zrc-sazu.si/a_beseda.html

British National Corpus (BNC); Brit Nemzeti Korpusz; http://www.natcorp.ox.ac.uk/

Brown University Standard Corpus of Present-Day American English; Brown Egyetem Mai Amerikai Angol Nyelvének Standard Korpusza, röviden Brown Korpusz. Kézikönyv: http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM

Brooklyn–Geneva–Amsterdam–Helsinki Parsed Corpus of Old English; Brooklyn–Geneva–Amsterdam–Helsinki Szintaktikailag Elemzett Óangol Korpusza. Ismertetés: http://www-users.york.ac.uk/~sp20/corpus.html

Cambridge International Corpus (CIC); Cambridge Nemzetközi Korpusz; http://uk.cambridge.org/elt/corpus/cic.htm

Cambridge and Nottingham Corpus of Discourse in English (CANCODE); http://www.cambridge.org/elt/corpus/cancode.htm

CHILDES Database; http://childes.psy.cmu.edu/

COBUILD (Collins Birmingham University International Language Databank); Kereső honlapja: http://www.collins.co.uk/Corpus/CorpusSearch.aspx

COBUILD Direct Corpus

COBUILD Spoken Corpus; COBUILD Beszélt Nyelvi Korpusza

Collins Word Web (a könyvben még ilyen néven nem szerepel); http://www.collins.co.uk/books.aspx?group=180

Corpus of English-Canadian Writing

Corpus du Théâtre religieux français du Moyen Âge; Középkori Francia Vallásos Színház Korpusza; http://www.byu.edu/~hurlbut/fmddp/

Corpus VALIFLOUI (Variétés Linguistiques du Français en Louisiane); http://languages.louisiana.edu/French/Valifloui.html

Cseh Nemzeti Korpusz; http://ucnk.ff.cuni.cz/

Dialogstrukturenkorpus

English-Swedish Parallel Corpus; Angol–svéd Párhuzamos Korpusz; http://www.englund.lu.se/content/view/66/127/

Eötvös Loránd Tudományegyetem Korpusza
EuroWordNet; http://www.illc.uva.nl/EuroWordNet/
FIDA (szlovén nyelv); http://www.fida.net/slo/index.html
FrameNet; http://framenet.icsi.berkeley.edu/
Francia Beszélt Nyelvi Korpusz
Freiburg Corpus; Freiburg–Brown (FROWN); (Freiburg–LOB [FLOB] Korpusz); Kézikönyv: http://khnt.hit.uib.no/icame/manuals/frown/INDEX.HTM
Freiburger Korpus; http://www.ids-mannheim.de/ksgd/agd/korpora/frkorpus.html
Hansard Corpus; http://www.isi.edu/natural-language/download/hansard/
Hong Kong University of Science and Technology (HKUST); Corpus of Learner English; Hongkongi Műszaki és Természettudományi Egyetem Angol Tanulói Korpusza
Hong Kong Corpus of Conversational English (HKCCE); Hongkongi Társalgási Angol Nyelv Korpusza; http://www.engl.polyu.edu.hk/Dept/Academic/Personal/ChengWinnie/HKCorpus_SpokenEnglish.htm
Horvát Nemzeti Korpusz; http://www.ffzg.hr/zzl/
Human Communication Research Centre's Map Task Corpus; Humán Kommunikációs Kutatási Központ Térkép Feladatának Korpusza; http://www.hcrc.ed.ac.uk/Site/MAPTASKD.html
Hunglish Korpusz; http://lab.mokk.bme.hu/eszkozok/hunglishkorpusz
International Computer Archive of Modern and Medieval English (ICAME); http://helmer.hit.uib.no/icame.html
International Corpus of English (ICE); Nemzetközi Angol Korpusz; http://www.ucl.ac.uk/english-usage/ice/
International Corpus of Learner English (ICLE); http://cecl.fltr.ucl.ac.be/research%20learner%20corpora.html
Japán Diákok Angol Nyelvű Korpuszai; http://leo.meikai.ac.jp/~tono/
JPU Corpus; Janus Pannonius Tudományegyetem Korpusza
Kanadai Francia Korpusz; http://www.spl.gouv.qc.ca/corpus/index.html
Kolhapur Corpus of Indian English (KOL), kézikönyv: http://khnt.hit.uib.no/icame/manuals/kolhapur/
Lancaster–IBM Spoken English Corpus; http://www.scs.leeds.ac.uk/nti-kbs/ai5/Misc/sec.html
Lancaster–Leeds Treebank
Lancaster–Oslo/Bergen Corpus (LOB)
Linguistic Data Consortium (LDC); Nyelvészeti Adatok Konzorciuma; http://www.ldc.upenn.edu/
London–Lund Corpus of Spoken English (LLC); London–Lund Beszélt Nyelvi Korpusz; kézikönyv: http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM
Longman Corpus Network (LCN); Longman Korpusz Hálózat; http://www.longman.com/dictionaries/corpus/lccont.html
1) Longman Learners' Corpus; Longman Nyelvtanulói Korpusz
2) Longman/Lancester English Language Corpus – LLELC; Longman/Lancester Angol Nyelvű Korpusza
3) Longman Spoken British Corpus; Longman Beszélt Nyelvi Koprusz, mely a BNC részét képezi
4) Longman Written American English; Longman Írott Nyelvi Amerikai Angol Nyelvi Korpusz
5) Longman Spoken American English; Longman Beszélt Nyelvi Amerikai Angol Nyelvi Korpusz; http://www.longman-elt.com/dictionaries/corpus/lccont.html
Magyar dalszövegek; http://www.recmusic.org/lieder/languages.html?LangId=14

Magyar Elektronikus Könyvtár; http://www.mek.iif.hu/porta/, új cím: http://mek.oszk.hu
Magyar Irodalmi és Köznyelv Nagyszótárának Korpusza / Magyar Történeti Korpusz; http://www.nytud.hu/hhc/
Magyar Nemzeti Szövegtár (MNSZ); http://corpus.nytud.hu/mnsz/bevezeto_hun.html
Magyar Webkorpusz; http://lab.mokk.bme.hu/eszkozok/webkorpusz/
negr@ korpusz; http://www.coli.uni-saarland.de/projects/sfb378/negra-corpus/negra-corpus.html
Német telefonbeszélgetések
Oxford Text Archive; Oxfordi Szövegarchívum; http://ota.ahds.ac.uk
PAROLE Francia Korpusz; http://www.elda.org/catalogue/en/text/W0020.html
Parsed Corpus of Early English Correspondence; A Szintaktikailag Elemzett Korai Angol Levelezés Korpusza
PELCRA (Polish and English Language Corpora for Research and Applications); http://www.uni.lodz.pl/pelcra/
Penn–Helsinki Parsed Corpus of Early Modern English; Penn–Helsinki Szintaktikailag Elemzett Korai Modern Angol Korpusz
Penn–Helsinki Parsed Corpus of Middle English; Penn–Helsinki Szintaktikailag Elemzett Közép Angol Korpusza; http://www.ling.upenn.edu/hist-corpora/
Pfeffer–Korpus; http://www.ids-mannheim.de/ksgd/agd/korpora/pfkorpus.html
Survey of English Usage (SEU); Angol Nyelvhasználati Felmérés; Intézet honlapja: http://www.ucl.ac.uk/english-usage/
Szerb Nyelv Korpusza; http://serbian-corpus.edu.yu/ie/eindex.htm
Szeged Korpusz; http://www.inf.u-szeged.hu/projectdirs/hlt/
Tiger Korpusz; http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERCorpus/
Tycho Brahe Parsed Corpus of Historical Portuguese;Tycho Brahe Történeti Portugál Korpusz; http://www.ime.usp.br/~tycho/corpus/
Wellington Corpus of Written New Zealand English; http://www.vuw.ac.nz/lals/corpora/index.aspx
WordNet; http://wordnet.princeton.edu/
World English Corpus; Világ Angol Korpusz [Macmillan Kiadó]
York–Helsinki Parsed Corpus of Old English Poetry; York–Helsinki Óangol Költészet Korpusza; http://www-users.york.ac.uk/~lang18/pcorpus.html
York–Toronto–Helsinki Parsed Corpus of Old English Prose; York–Toronto–Helsinki Szintaktikailag Elemzett Óangol Prózai Korpusz; http://www-users.york.ac.uk/~lang22/YcoeHome1.htm

KORPUSZNYELVÉSZETI ALAPFOGALMAK

(magyar, angol, francia, német és japán nyelven, magyar meghatározásokkal)[33]

M: **Adatbázis**
A: database
F: base de données (f)
J: データベース
Strukturált / rendezett információt tartalmazó egység (pl. telefonkönyv, könyvtári katalógus), mely lehetővé teszi az adatok kívánt szempontok szerinti gyors lekérdezését.

M: **adatvezérelt nyelvtanulás**
A: data-driven language learning (DDL)
F: apprentissage de langue dirigé (induit) par les données linguistiques
N:
J: データ駆動型学習
Tim Johns, a Birminghami Egyetem tanára alkotta a kifejezést. Az autentikus nyelvi példákra épülő, gyakran konkordanciákat felhasználó nyelvtanulásra és tanításra használt kifejezés.

M: **ágbank** / szintaktikai adatbázis
A: treebank / syntactic database
F: banque d'arbres / base de données syntaxiques
N: Treebank (f) / syntaktisch analysiertes Korpus
J: 構文解析コーパス
Szintaktikailag elemzett mondatok adatbázisa.

M: **ágrajz**
A: tree diagram
F: arbre (m)
N: Baumdiagramm (n)
J: 樹形図
A szintaktikai elemzések vizuális megjelenítése, mely fára emlékeztet.

M: **általános korpusz**
A: general corpus
F: corpus de référence (m)
N: allgemeines Korpus
J: 汎用コーパス
A teljes nyelvhasználat vizsgálata céljából készült, így elsősorban lexikográfiai vizsgálatokra használják.

M: **annotáció**
A: annotation
F: annotation (f)
N: Annotation (f), Annotierung (f)
J: １）テキスト上に付帯情報をつけること　２）情報付与
Olyan információ, mely az eredeti szövegben nem szerepel, hanem a szövegfeldolgozás során kerül a szövegbe. A szövegre vonatkozik, de a szövegtől egyértelműen megkülönböztethető. Az annotáció leggyakoribb formái a címkézés és a szintaktikai elemzés, de a nyelvtanulói korpuszok hibakódjai is idetartoznak.

M: **annotált korpusz**
A: annotated corpus
F: corpus annoté / étiqueté (m)
N: annotiertes Korpus
J: 情報付与コーパス
A szövegre és egységeire (bekezdés, szavak stb.) vonatkozó információval ellátott korpusz.

M: **átírás**
A: transcription
F: transcription (f)

[33] A fogalomtár a korpusznyelvészetben használt alapfogalmakat tartalmazza. A nyelvtudomány más területeiről már ismert nyelvészeti fogalmak itt nem szerepelnek, hiszen azokat más magyar nyelvű könyvekben is meg lehet találni, pl. *Nyelvi fogalmak kisszótára* (Kugler & Tolcsvai Nagy 2000) vagy *A nyelv enciklopédiája* (Crystal 2003). A fogalomtárat folyamatosan bővítjük a www.korpusz.com lapon, így a korpusz@hotmail.com címre küldött javaslataikkal és észrevételeikkel segíthetik ennek fejlesztését és pontosítását.

N: Transkription (f)
J: 転写
A beszélt nyelvi szövegek lejegyzése, mely több-kevesebb részletességgel történhet. A legegyszerűbb formája a puszta szavak átírása, de legtöbbször az elhangzott szavak mellett az élőbeszéd lehető legtöbb jellemzőjét igyekszik megragadni, pl. a szünetet, hangerősséget, hanglejtést, hangsúlyt stb.

M: **automatikus szintaktikai elemzés**
A: parsing
F: analyse syntaxique automatique – parsage (kanadai francia nyelvből)
N: Parsing
J: 構文解説; 構文解析
Az annotáció egy fajtája, melynek során a szöveget és a mondatokat szintaktikai egységekre bontják.

M: **automatikus szintaktikai elemző program**
A: parser
F: parseur (m)
N: Parser (m)
J: 構文解説プログラム; 構文解析ツール
A szintaktikai elemzést végző számítógépes program.

M: **beszédfelismerés**
A: voice recognition
F: reconnaissance (f) vocale
N: Spracherkennung (f)
J: 音声認識
Az a technológia, amely a hangot felismeri és automatikusan szöveggé alakítja.

M: **címke**
A: tag
F: étiquette (f)
N: Tag (igenévként: getaggtes Korpus)
J: 言語的標識
Az annotáció során a szövegegységhez kapcsolódó és azt jellemző kód.

M: **címkézés**
A: tagging
F: étiqueter
N: (Wortarten-)Tagging (n), Zuordnung (f) von Tags
J: 標識付け
Az annotáció egy fajtája, amikor a szöveg egységeit, legtöbbször a szavakat, egy vagy több, az adott egységre vonatkozó információt tartalmazó megjegyzéssel látják el. Legtöbbször a szófaji címkézést értik alatta.

M: **címkéző program**
A: tagger
F: étiqueteur (m)
N: Tagger (m), -n (plur.)
J: タグ付けプログラム
A címkézést automatikusan végző program.

M: **COCOA utalások**
A: COCOA reference
F: référence de COCOA
N:
J: COCOA　形式
Olyan hegyes zárójel-pár (<>), amely egy információkódot és annak egy valós megnyilvánulását tartalmazza. Pl. ha C=cím, akkor <C Egri csillagok>.

M: **csontváz elemzés** (szintaktikai)
A: skeleton / shallow parsing
F: parsage (m) squelettique
N: skeleton parsing, flache Analyse
J: 簡略解析
Részleges szintaktikai elemzés, amelyben egyes csoportokat további részelemekre lehetne bontani.

M: **dinamikus korpusz**
A: dynamic corpus
F: corpus dynamique
N: dynamisches Korpus
J: 動的コーパス
Folyamatosan bővülő korpusz, ahol a szövegtípusok aránya nem állandó.

M: **előfordulás**
A: occurrence
F: occurrence (f)
N: Vorkommen (n)
J: 出現
Egy adott szóalak adott korpuszban való előfordulásának a száma

M: **fordítási ekvivalencia / megfelelés**
A: translation equivalence
F: équivalence (f) de traduction

N: Entsprechung (einer Übersetzung)
J: 等価
A különböző fordításelméletek különbözőképpen határozzák meg e kulcsfogalmat. *A* nyelvről *B* nyelvre való fordításkor a fordítási megfelelőket a hétköznapi életben a kétnyelvű szótárakban keressük, de a szövegek közötti megfelelésnek a diskurzus szintjén kell létrejönni. A gépi fordítások „furcsaságai" bizonyítják ezt a legszemléletesebben. A párhuzamos korpusz esetében sem szavakat, hanem jelentésegységeket jelölnek meg fordítási megfelelőként.

M: **fordítási korpusz**
A: translation corpus
F: corpus de traduction
N: Übersetzungskorpus (n)
J: 翻訳コーパス
Kizárólag fordítás eredményeként létrejött szövegeket tartalmaz.

M: **Hapax legomenon / legomena**
A: Hapax legomenon, pl. legomena
F: hapax (m)
N: Hapaxlegomenon (n)
J: ハパックス
A korpuszban egyszer előforduló szóalak.

M: **homográf**
A: homograph
F: homographe (m)
N: Homograph (m)
J: 同綴異議語
Azonos írásképű, de különböző jelentésű szó.

M: **idióma elv**
A: idiom principle
F: principe (m) d'idiome
N:
J: 慣用原則
Sinclair nyelvelméleteben az az elv, mely szerint a jelentés szóegyüttesekből és nem különálló szavak együttes jelentéséből származik, és ezek a memóriában is így tárolódnak, vö. szabad választás elve.

M: **jelentés egyértelműsítés**
A: word sense disambiguation
F: désambiguïsation lexicale
N: Wortsinndisambiguierung (f)
J: 曖昧性除去、両義性除去
A szövegben szereplő azonos alakú, de különböző jelentésű szavak adott esetben érvényes jelentésének meghatározása. (pl. a *vár* szó főnévi vagy igei jelentése)

M: **kiegyensúlyozott korpusz**
A: balanced corpus
F: corpus équilibré
N: balanciertes Korpus
J:
Olyan korpusz, amelyben az egyes szövegtípusok és azok mennyisége többé-kevésbé hűen tükrözik a valóságban betöltött arányokat. Az általános nyelv leírásának céljából készül, így általában hatalmas mennyiségű szöveget foglal magába, melyeket elsősorban lexikográfiai célokra használnak. Lásd: általános korpusz

M: **klón**
A: clone
F: clone (m)
N: Klon (m)
J: クローン
Egy bizonyos korpusz szerkezetét követő korpusz. Legtöbbször a Brown Korpusz mintájára készült korpuszokra utal e kifejezés, de bármely más korpusz mintáját követőre is alkalmazható.

M: **kolligáció**
A: colligation
F: colligation (f)
N: Kolligation (f)
J: コリゲーション,
単語と品詞との共起パターン
Bizonyos szavak bizonyos nyelvtani szerkezetben való előfordulása, mely a valószínűségen alapuló várható véletlen együttes előfordulásnál magasabb, és sok esetben előre kitalálható, pl. *elhatározta, hogy*; vö. kollokáció

M: **kollokáció**
A: collocation
F: collocation (f)
N: Kollokation (f)
J: 連語、コロケーション,
単語の共起パターン
Bizonyos szavak együttes előfordulása, mely a valószínűségen alapuló várható véletlen együt-

tes előfordulásnál magasabb, és sok esetben előre kitalálható, pl. *szőke haj*; vö. kolligáció

M: **konkordancia**
A: concordance
F: concordance (f) ou ligne de contexte
N: Konkordanz (e)
J: コンコーダンス
Egy adott szó vagy kifejezés szövegben szereplő összes előfordulását szövegkörnyezetében bemutató lista.

M: **konkordanciaprogram**
A: concordancer
F: logiciel de concordances (m)
N:
J: コンコーダンサー / コンコーダンス･プログラム
Konkordanciákat létrehozó program.

M: **korpusz**
A: **corpus pl. corpora**
F: corpus (m)
N: Korpus (n), Korpora
J: コーパス
Nyelvészeti vizsgálatok céljából, bizonyos szempontok alapján összeválogatott írott vagy beszélt nyelvi szövegek gyűjteménye.

M: **korpusz alapú**
A: corpus based (adj)
F: basé, -e sur le coprus
N: korpusanalytisch basiert
J: コーパスに基づく
Bizonyos problémák vizsgálatára a korpuszelemzés eredményei adnak választ, a nyelvészeti vizsgálatok a korpuszelemzésre épülnek.

M: **korpusz informált**
A: corpus informed
F: informé, - e par le corpus
N:
J:
Nem kizárólagosan korpuszelemzésre épülő kutatás. A korpuszvizsgálatok eredményeit csak megerősítésként igénylik.

M: **korpusznyelvészet**
A: corpus linguistics
F: linguistique (f) de corpus
N: Korpuslinguistik (f)
J: コーパス言語学
Azon nyelvészeti irányzat, mely a nyelv és nyelvhasználat vizsgálatát speciális módszerek és számítógépes programok segítségével korpuszra alapozva végzi.

M: **korpuszvezérelt**
A: corpus driven
F: conduit -e par le corpus
N:
J: コーパス駆動的
Az empirikus kutatásnak az a változata, amely a korpusz elemzése folyamán felfedezett szabályszerűségek alapján von le következtetéseket.

M: **KWAL formátum**
A: Key word and line (KWAL)
F: mot clé et ligne en contexte
N:
J: キーワードとコンコーダンス･ラインを中心に前後に文脈を表示する
E formátumban a keresett szót tartalmazó sor több sort kitevő szövegkörnyezetével együtt látható.

M: **KWIC formátum**
A: key word in context KWIC
F: mot clé en contexte
N:
J: キーワードを中心に前後に文脈を表示する
E formátumban a keresett szó a szövegkörnyezetével együtt látható, mely általában a számítógép monitorán egy sort tesz ki. A keresett szó általában középen helyezkedik el.

M: **lemma**
A: lemma
F: lemme
N: Lemma (n)
J: 見出し語
Az azonos szótőből származó összes (általában azonos szófajú) szóalakot átfogó kategória, pl. *ugrál, ugrik, ugrott* stb. A kutatás igényeihez igazodva különböző szófajú alakok is tartozhatnak egy lemmába.

M: **lemmatizálás**
A: lemmatization
F: lemmatisation (f)

N: Lemmatisierung (f)
J: 見出し語化
A különböző szóalakok lemmákba való csoportosítása.

M: **lemmatizáló program**
A: lemmatizer
F: logiciel de lemmatisation
N: Lemmatiser (m)
J:
A különböző szóalakok lemmákba való csoportosítását automatikusan végző program.

M: **lexika**
A: lexis
F: lexique (f)
N: Lexika (plur.)
J: 語彙目録 語彙
Egy nyelv szókészlete, általában a nyelvtannal szembeállítva használatos.

M: **lexikon**
A: lexicon
F: léxicon (m) lexique (m)
N: Lexikon (n), Wortbestand (m)
J: 用語集
Jelentése gyakorlatilag szótár vagy szókincs, de a számítógépes programok szókincs adatbázisaira is legtöbbször ezt a kifejezést használják.

M: **mesterséges intelligencia**
A: AI artificial intelligence
F: Intelligence (f) Artificielle
N: künstliche Intelligenz
J: 人工知能、AI
Az emberi agy működését vizsgálja azzal a céllal, hogy számítógépek segítségével szimulálja azt.

M: **MI együttható**
A: MI score (mutual information)
F: score (m) d'information mutuelle
N:
J: (特定の２語間の連想関係の強さを計る尺度)
Két szó **tényleges** együttes előfordulását a valószínűségi számítások alapján **„várható"** előfordulással való összehasonlítás eredménye (informatikai elméletből származik). A kollokációk vizsgálatánál használják.

M: **mintavétel**
A: sampling
F: échantillonnage (m)
N:
J: 標本抽出 サンプリング
Szövegminták kiválasztása adott populációra (szövegtípusra, regiszterre) vonatkozó vizsgálatok végzése céljából.

M: **monitor korpusz**
A: monitor corpus
F: corpus de suivi (de baromètre)
N: Monitorkorpus (n)
J: モニター・コーパス
Olyan korpusz, amely lehetővé teszi a nyelv rövidtávú változásának elemzését a szövegek korpuszon belüli elkülönítésével. Rendszeresen fejlesztik a korpuszt, de az arányokat mindig megtartják.

M: **n-gram**
A: n-gram
F: n-gramme (f)
N:
J: (文字列の連鎖を集計する機能)
Az n-gramok a szövegben szereplő több egységből álló szerkezetek/szövegsablonok korpuszvezérelt elemzését segítik elő. Az *n* helyére kerülő szám határozza meg, hogy a szövegben szereplő szavakat hányas egységekbe csoportosítjuk. Pl. *A macska az asztalon ül.* 3-gram esetében: *a macska az, macska az asztalon,* és *az asztalon ül* egységekre bonthatók.

M: **nyelvtanulói korpusz**
A: learner corpus
F: corpus d'étudiants de langue
N:
J: 学習者コーパス
Nem anyanyelvi beszélők nyelvi megnyilvánulásaiból összeállított korpusz.

M: **összehasonlítható korpusz**
A: comparable corpus
F: corpus comparable
N:
J: 比較コーパス・コンパラブルコーパス (多言語間または複数言語変種間の比較ができるように、同じコーパスデザインで編纂されたコーパス)

Két vagy több olyan korpusz, amelynek szerkezete és mérete hasonló (pl. angol és magyar nyelvű üzleti levelezés).

M: **párhuzamos korpusz**
A: parallel corpus
F: corpus parallèle
N: Parallelkorpus (n)
J: パラレル・コーパス (2カ国語以上による同じ内容のコーパス)
Olyan korpusz, amely több mint egy nyelven tartalmazza ugyanazt a szöveget vagy szövegeket.

M: **pedagógiai korpusz**
A: pedagogic corpus
F: corpus pédagogique
N: pädagogisches Korpus
J: 教育型コーパス
Dave Willis szóhasználatában azon szövegek összessége, amellyel a nyelvtanulók egy kurzus alatt találkoznak.

M: **példány**
A: token
F: occurrence (f)
N: Token (n)
J: トークン(総語数)
Egy szövegben akár többször is előforduló bármely szó; vö. szövegszó.

M: **probabilisztikus módszerek**
A: probabilistic methods
F: méthodes probabilistes (f)
N: Wahrscheinlichkeitsmethoden
J: 確率論的手法
A statisztikai valószínűségen alapuló módszerek.

M: **reguláris kifejezés (regexp)**
A: regular expression
F: expression (f) régulière
N: regulärer Ausdruck (m)
J: 正規表現
A programozásban használt olyan kifejezés, amelyet főleg szűrőknél, minták feldolgozására és keresésére használnak.

M: **reprezentatív**
A: representative
F: représentatif, -ive
N: repräsentativ
J: 代表的 典型的な
Olyan minta, amely a populációra jellemző jegyek összességét a lehető legnagyobb mértékben megközelíti.

M: **SGML szabványos, általános leíró nyelv**
A: SGML Standard Generalized Markup Language
F: language standard de balisage SGML
N:
J: 形式
Szöveges állományok belső szerkezetének (fejezetek, bekezdések, lábjegyzetek stb.) jelölésére használható szabvány.

M: **statikus korpusz**
A: static corpus
F: corpus statique
N: statisches Korpus
J:
Olyan korpusz, amelynek tartalma nem változik.

M: **statisztikai jelentősség**
A: significance
F: significance (f)
N: Signifikanz (f)
J: 有意水準 (level)
Nem a véletlenen múló statisztikai eredmény, mely alapján következtetések vonhatók le.

M: **szabad választás elve**
A: open choice principle
F: principe (m) de choix
N:
J:
Sinclair szóhasználatában az idióma elvvel ellentétes elv, mely szerint a jelentés az egyes szavak jelentésének összességéből jön létre, így minden szó után szabadon választható meg a következő, vö. idióma elv.

M: **számítógépes nyelvészet**
A: computational linguistics
F: linguistique informatisée?
N: Komputerlinguistik (f)
J: コンピュータ的言語学
A nyelv vizsgálatához számítástechnikai elveket és módszereket használó tudományterület.

M: **szemantikai prozódia**
A: semantic prosody
F: prosodie sémantique
N: semantische Prosodie
J: 意味的プロソディ
Bizonyos szavak csak bizonyos jelentéstartalmú szavakkal vagy nyelvi szerkezetekkel együtt fordulnak elő. Pl. az ***okoz*** általában negatív dolgokkal: *bánat, baleset, kár* stb.

M: **szóalak/típus**
A: type
F: type (m)
N: Typ (m)
J: タイプ(異なり語数)
A szövegben előforduló különböző írásképű szó (pl. bot, botot).

M: **szókincstár (thesaurus)**
A: thesaurus
F: thésaurus (m)
N: Thesaurus (m)
J: 類語辞典, シソーラス、語彙分類集
Olyan szótár, amelyben a szavakat jelentésük alapján csoportosítják, nem pedig ábécérendben.

M: **szöveggyűjtemény /szövegarchívum**
A: text collection/text archive
F: collection de textes/archive de textes
N: Textarchiv (n)
J: テキスト集合体
Szövegek esetleges gyűjteménye, tárháza; vö. korpusz.

M: **szövegkörnyezet**
A: context
F: contexte (m)
N: Kontext (m)
J: 文脈,前後
A korpusznyelvészetben használt meghatározás szerint egy szót vagy kifejezést közvetlen megelőző és követő szövegrészlet. Ennek segítségével lehet az adott szó vagy kifejezés jelentését egyértelműsíteni.

M: **szövegszinkronizálás**
A: alignment
F: alignement (m)
N: Alignierung (f)
J:
Két szöveg/korpusz egymáshoz való igazítása, melynek során az összetartozó elemeket megjelölik. Így az egyikben történő keresés-kor a másikban hozzárendelt adatok is megjelennek. Fordítási vagy párhuzamos korpuszok esetében igen gyakori.

M: **szövegszó**
A: running words
F: mot (graphique)
N: laufende Wörter/Wortformen
J:
Olyan betűcsoport, amelyet mindkét oldalon szóköz választ el. Esetenként a jobb oldali szóközt írásjel előzi meg.

M: **T-együttható, Tí-szkór**
A: T-score
F: score T de cooccurrence
N:
J: T-score
(特定の２語間にj何らかの連想関係があることを主張することができる確信度を計る尺度)
Két elem közötti összefüggésre vonatkozó, statisztikai együttható, a kollokációk vizsgálatánál használják.

M: **TEI (szövegkódolási ajánlás)**
A: (TEI) Text Encoding Initiative

F: TEI (f) Initiative de documentation de textes

N:
J: テキスト電子化の規格化
Olyan nemzetközi és interdiszciplináris szabvány, amely a szövekódolást igyekszik egységessé tenni. A pontosságra és egyszerűségre törekednek. A szabvány fejlesztésére 1987-ben konzorciumot hoztak létre.

M: **teljes szintaktikai elemzés**
A: full parsing
F: analyse (f) complète
N:
J: 詳細解析
A mondat minden egységének legkisebb szintaktikai egységre való lebontása.

M: **Természetes nyelvfeldolgozás**
A: NLP Natural Language Processing

F: traitement automatique du language naturel (m) (TALN)
N:
J: 自然言語処理
A nem formális, azaz emberi nyelvek számítógépes feldolgozása.

M: **többnyelvű korpusz**
A: multilingual corpus
F: corpus multillingue
N: mehrsprachiges Korpus
J: 多言語コーパス
Több nyelven tartalmaz szövegeket. Ezek típusuk szerint különbözők lehetnek: fordítási, összehasonlítható vagy párhuzamos korpusz.

M: **többváltozós statisztikai elemzés**
A: multivariate statistics
F: analyse (f) statistique multivariée
N: multivariate Statistik
J: 多変量統計
Olyan statisztikai elemzések, amelyek egyszerre több változó közötti kapcsolatokat vizsgálnak.

M: **történeti/ diakronikus korpusz**
A: historical/diachronic corpus
F: corpus diachronique
N: historisches/diakronisches Korpus
J: 通時コーパス
A nyelv történeti változásának tanulmányozása céljából nem kortárs szövegeket tartalmazó korpusz.

M: **tudásbázis**
A: knowledge base
F: base de connaissance (f)
N: Knowledge-Base (f)
J:
Egy mesterséges intelligencia program műveletek elvégzéséhez szükséges szabályainak forrása, melyek formális nyelven íródnak.

M: **tudatosság javítása / tudatosságot javító**
A: awareness/consciousness raising
F: élévation (f) du niveau de conscience
N: Bewusstseinerhebung (f)
J: 気づき / 意識化
A nyelvtanítás azon elve, mely szerint a nyelvtan explicit tanítása helyett a nyelvtanuló nyelvi/nyelvtani tudatát olyan feladatokkal emeli, mely a diákok részéről aktív megfigyelést és következtetések levonását igényli.

M: **vizsgált csomópont/adat (nód)**
A: node
F: noeud (m)
N:
J: 中心点
A konkordanciákban és az elemzésekben a vizsgált adat helyett gyakran megjelenő kifejezés.

M: **Z-együttható, Zí-szkór**
A: Z-score
F: score Z de cooccurrence
N:
J: Z値
Két elem közötti összefüggésre vonatkozó, statisztikai együttható, a kollokációk vizsgálatánál használják.

BIBLIOGRÁFIA

A

A könyv írásakor a szerző, Szirmai Monika, az alábbi munkáit is felhasználta, de ezekre külön nem hivatkozott:

Disszertációk

1. (1994). Translation equivalence between English and Hungarian (kiadatlan MA disszertáció, témavezető: John Sinclair) University of Birmingham
2. (2001). *The theory and practice of corpus linguistics*. Debrecen: Debreceni Egyetem Kossuth Egyetemi Kiadója (Doktori Értekezések Vol. 12, témavezető: Hollósy Béla)

Szakcikkek, könyvrészletek

1. (1996). Corpus linguistics in education. *The Journal of Kanda University of International Studies, No. 8*, 273–287.
2. (1997). A (corpus) linguistic theory of translation. *The Journal of Kanda University of International Studies, No. 9*, 269–280.
3. (1998). Corpus linguistics: An introduction. *The Journal of Kanda University of International Studies, No. 10*, 381–395. Reprinted in: (2002) 「英語学論説資料収録論文—覧」第34号 (2000年分) (pp. 14–21).
4. (2001). Corpus tools in language teaching and learning. In J. White (Ed.), *Fleat IV. (The 4th Conference on Foreign Language Education and Technology)* (pp. 146–151 [On CD-ROM]). Kobe, Japan: the Japan Association for Language Education and Technology.
5. (2002). A brief history of corpora [Proceedings] In: S. Petersen & M. Kruse (Szerk.) The changing face of CALL: Emerging technologies, emerging pedagogies (pp. 87–91). Nagoya: JALT CALL SIG.
6. (2002). Corpus linguistics in Japan: Its status and role in education (pp. 91–108). In *The changing face of CALL: A Japanese perspective.* Lisse, Hollandia: Swets & Zeitlinger Publishers (now Taylor & Francis).

Konferencián és tudományos fórumon elhangzott előadások

1. (1997). Classroom applications of concordancing programs. *1st Pan-Asia and 17th Thai TESOL International Conference*. Bangkok, Thailand.
2. (1997). Students as researchers, *The 23rd International Conference on Language Teaching/Learning*. Hamamatsu, Japan.
3. (1997). Teaching collocations, *The 23rd International Conference on Language Teaching/ Learning*. Hamamatsu, Japan.
4. (1998). Concordancing: A powerful tool for teachers and students. *The Seventh International Symposium on English Teaching*. Taipei, Taiwan.

5. (2000). Make your own "Contexts" [Workshop] *JALTCALL*, Hachioji, Japán.
6. (2002). A Rationale for Corpus Linguistics. *237th Meeting of the Hiroshima Language and Culture Circle,* Hiroshima Women's University, Hiroshima.
7. (2004). Corpus methods in education. *1st Corpus Meeting of Hiroshima International University,* HIU Education Center, Hiroshima.

B

Aarts, B. (2000). Corpus linguistics, Chomsky and fuzzy tree fragments. In C. Mair & M. Hundt (Szerk.), *Corpus linguistics and linguistic theory* (5–13). Amsterdam–Atlanta, GA: Rodopi.

Aarts, J., & Meijs, W. (Szerk). (1984). *Corpus linguistics: Recent developments in the use of computer corpora in English language research*. Amsterdam: Rodopi.

Altenberg, B., & Aijmer, K. (2000). The English-Swedish Parallel Corpus: A resource for contrastive reserach and translation studies. In C. Mair & M. Hundt (Szerk.), *Corpus linguistics and linguistic theory* (15–33). Amsterdam – Atlanta, GA: Rodopi.

Andor, J. (2004). The master and his performance: An interview with Noam Chomsky. *Intercultural Pragmatics, 1* (1), 93–111.

Anthony, L. (2004). AntConc (Version 2.6.0).

Aston, G. (Ed.). (2001). *Learning with corpora*. Houston: Athelstan.

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing, 7* (1), 1–16.

Atwell, E. (1993). Corpus-based statistical modelling of English grammar. In C. Souter & E. Atwell (Szerk.), *Corpus-based computational linguitics* (195–214). Amsterdam: Rodopi.

Atwell, E., Leech, G., & Garside, R. (1984). Analysis of the LOB corpus: Progress and prospects. In J. Aarts & W. Meijs (Szerk.), *Corpus linguistics: Recent developments in the use of computer corpora in English language research* (41–52). Amsterdam: Rodopi.

Bach, I. (1995). *Számítástechnikai nyelvészet*. Budapest: Budapesti Műszaki Egyetem.

Barlow, M. (1999). MonoConc 1.5 and ParaConc (Sotware assessment). *International Journal of Corpus Linguistics, 4* (1), 173–184.

Bauer, L. (1993a). *Manual of Information to Accompany the Wellington Corpus of New Zealand English*. Wellington: Department of Linguistics, Victoria University of Wellington.

Bauer, L. (1993b). Progress with a corpus of New Zealand English and some early results. In C. Souter & E. Atwell (Szerk.), *Corpus-based computational linguitics* (1–10). Amsterdam: Rodopi.

Benson, M. & Benson, E. (1993). *Russian-English dictionary of verbal collocations*. Amsterdam–Philadelphia: John Benjamins.

Benson, M., Benson, E., & Ilson, R. (1986). *The BBI combinatory dictionary of English: a guide to word combinations*. Amsterdam: John Benjamins Publishing Company.

Berry, R., & Cobuild. (1993). *Collins Cobuild English guides 3: Articles*. London: Harper-Collins Publishers.

Berry, R., & Cobuild. (1997). *Collins Cobuild English guides 10: Determiners & quantifiers*. London: HarperCollins.

Biber, D. (1990). Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing, 5*, 257–269.

Biber, D., Conrad, S., & Leech Geoffrey, N. (2002). *Longman Student Grammar of Spoken and Written English*. London: Longman.

Biber, D., & Finegan, E. (1994). Intra-textual variation within medical research articles. In Oostdijk & d. Haan (Szerk.), *Corpus-based reasearch into language: In honour of Jan Aarts* (201–222). Amsterdam: Rodopi.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* London: Longman.

Bottyán Gergely (2005). *A korpuszhasználatban rejlő kiaknázatlan lehetőségekről*. A XV. Magyar Alkalmazott Nyelvészeti kongresszuson elhangzott előadás, (megjelenés alatt).

Brigham Young, U. (1989). *WordCruncher*. Provo, Utah: Electronic Text Corporation.

Burnard, L. (1992). The Text Incoding Initiative: A progress report. In G. Leitner (Szerk.), *New directions in English language corpora: Methodology, results, software developments* (97–107). Berlin: Mouton de Gruyter.

Burnard, L. (1995). The Text Encoding Initiative: An overview. In G. Leech, G. Myers & J. Thomas (Szerk.), *Spoken English on computer: Transcription, mark-up and application* (69–81) London: Longman.

Burnard, L. (2000). *Where did we go wrong? A retrospective look at the design of the BNC*. Paper presented at the Fourth International Conference on Teaching and Language Corpora (TALC), Graz.

Burnard, L., & McEnery, T. (Szerk.). (2000). *Rethinking language pedagogy from a corpus perspective* (Vol. 2). Frankfurt am Main: Peter Lang GmbH.

Bussmann, H. (1996). *Routledge dictionary of language and linguistics* (G. Trauth & K. Kazzazzi, Ford.). London & New York: Routledge.

Cambridge learner's dictionary (Semi-bilingual version). (2nd ed.) (2004). Tokyo: Cambridge University Press–Shogakukan Inc.

Capel, A. (1993). *Prepositions*. London: HarperCollins.

Carpenter, E., & Cobuild. (1993). *Collins Cobuild English guides 4: Confusable words*. London: HarperCollins Publishers.

Chalker, S., & Cobuild. (1996). *Collins Cobuild English guides 9: Linking words*. London: HarperCollins.

Cheepen, C. (1995). Discourse considerations in transcription and analysis. In G. Leech, G. Myers & J. Thomas (Szerk.), *Spoken English on computer: Transcription, mark-up and application* (135–143): Longman.

Cheng, W., & Warren, M. (1999). Facilitating a description of intercultural conversations: The Hong Kong Corpus of Conversational English. *ICAME Journal* (23), 5–20.

Chomsky, N. (1957). *Syntactic stuctures*. The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, Mass: The MIT Press.

Clear, J. (1992). Corpus sampling. In G. Leitner (Szerk.), *New directions in English language corpora: Methodology, results, software developments* (21–32). Berlin: Mouton de Gruyter.

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM, 13* (6), 377–387.

Cook, G. (1995). Theoretical issues: Transcribing the untranscribable. In G. Leech, G. Myers & J. Thomas (Szerk.), *Spoken English on computer: Transcription, mark-up and application* (35–54): Longman.

Cowie, A. P. (1999). *English Dictionaries for Foreign Leraners*. Oxford: Clarendon Press.

Crowdy, S. (1995). The BNC spoken corpus. In G. Leech, G. Myers & J. Thomas (Szerk.), *Spoken English on computer: Transcription, mark-up and application* (224–234) London: Longman.

Crystal, D. (1992). *An encyclopedic dictionary of language and languages*. Harmondsworth: Penguin Books.

Crystal, D. (2003). *A nyelv enciklopédiája* (László Zs., Rebrus P., Szemere P., Szentgyörgyi S., Szentgyörgyi-Kiss K., Szűcs T., Vinkler Zs. & Zólyomi G. Ford.). Budapest: Osiris.

Deignan, A., & Cobuild. (1995). *Collins Cobuild English guides 7: Metaphor*. London– Birmingham: HarperCollins–University of Birmingham.

Dodd, B. (Ed.). (2000). *Working with German corpora*. Birmingham: University of Birmingham Press.

Edwards, J. A. (1995). Principles and alternative systems in the transcription, coding and mark-up of spoken discourse. In G. Leech, G. Myers & J. Thomas (Szerk.), *Spoken English on computer: Transcription, mark-up and application* (19–34) London: Longman.

Encyclopedia Americana. (1999). Chicago: Grolier Educational (Version 5.0).

Fang, A. C. (1992). Building a corpus of computer science English. In J. Aarts, P. de Hann & N. Oostdijk (Szerk.), *English language corpora: Design, analysis and exploitation* (73–78). Amsterdam: Rodopi.

Fellbaum, C. (Szerk.). (1998). *WordNet: An electronic lexical database*. Cambridge, Mass: The MIT Press.

Fillmore, C. J. (1992). "Corpus linguistics" or "computer-aided armchair linguistics". In J. Svartvik (Szerk.), *Directions in corpus linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991* (35–60). Berlin–New York: Mouton de Gruyter.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955. *Studies in Linguistic Analysis, Special Volume, Philological Society*, 1–32.

Firth, J. R. (1957). Modes of meaning. in *Papers in linguistics 1934–1951*. London: Oxford University Press, 190–215.

Francis, W. N. (1982). Problems of assembling and computerizing large corpora. In S. Johansson (Szerk.), *Computer corpora in English language research* (7–24.). Bergen: Norwegian Computing Centre for the Humanities.

Francis, W. N., & Kučera, H. (1964). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin Company.

Fromkin, V. (1968). Speculations on Performance Models. *Journal of Linguistics, 4*, 47–68.

Garside, R. (1995). Grammatical tagging of the spoken part of the British National Corpus: A progress report. In G. Leech, G. Myers & J. Thomas (Szerk.), *Spoken English on computer: Transcription, mark-up and application* (161–167) London: Longman.

Ghadessy, M., Henry, A., & Roseberry, R. L. (Szerk.). (2001). *Small Corpus Studies and ELT: Theory and Practice.* Amsterdam: John Benjamins.

Goldstine, H. H. (1993). *The computer from Pascal to von Neumann*. Princeton, N. J.: Princeton University Press.

Goodale, M. (1995a). *Phrasal verbs*. London: HarperCollins.

Goodale, M. (1995b). *Tenses*. London: HarperCollins.

Granger, S. (1993). International Corpus of Learner English. In M. G. Aarts Jan, d. Haan Pieter & N. Oostdijk (Szerk.), *English language corpora: Design, analysis and exploitation: Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora, Nijmegen 1992* (57–72). Amsterdam; Atlanta, GA: Editions Rodopi B.V.

Granger, S. (1994). The learner corpus: A revolution in applied linguistics. *English Today, 10* (3), 25–29.

Granger, S. (1996). Learner English around the world. In S. Greenbaum (Szerk.), *Comparing English world-wide* (13–24). Oxford: Clarendon Press.

Granger, S. (Szerk.). (1998). *Learner English on computer*. London: Longman.

Granger, S., Hung, J., & Petch-Tyson, S. (Szerk.). (2002). *Computer learner corpora, second language aquisition and foreign language teaching* (Vol. 6). Amsterdam: John Benjamins Publishing Company.

Greaves, C. (2003). ConcApp. Public Version.

Greenbaum, S. (1992). A new corpus of English: ICE. In J. Svartvik (Szerk.), *Directions in corpus linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991* (pp. 171–179). Berlin–New York: Mouton de Gruyter.

Habert, B. (1999). Un corpus clé pour le français actuel. Letöltve: 2004. november 27.

Habert, B., Nazarenko, A., & Salem, A. (1997). *Les linguistiques de corpus*. Paris: Armand Colin/ Masson.

Halliday, M. A. K. (1978). *Language as social semiotic*. Glasgow: Edward Arnold.

Haslerud, V., & Stenström, A.-B. (1995). The Bergen Corpus of London Teenager Language (COLT). In G. Leech, G. Myers & J. Thomas (Szerk.), *Spoken English on computer: Transcription, mark-up and application* (235–242) London: Longman.

Hornby, A. S., Gatenby, A. V., & Wakefield, H. (1942). *Idiomatic and Syntactic English Dictionary*. Tokyo: Kaitakusha.

Hornby, A. S., Gatenby, A. V., & Wakefield, H. (1948). *A learner's dictionary of current English*. Oxford: Oxford University Press.

Horváth, J. (2001). *Advanced writing in English as a foreign language: A corpus-based study of processes and products*. Pécs: Lingua Franca Csoport.

Horváth, J. (1999). Collins COBUILD home page. *Modern Nyelvoktatás, 5* (1), 83–84.

Horváth, J. (2000). A JPU korpusz kialakítása és tartalma. *Modern Nyelvoktatás, VI* (2–3), 42–49.

Horváth, J. (2002). Pedagogical annotation of learner corpora. In B. Hollósy & J. Kiss-Gulyás (Szerk.), *Studies in Linguistics Volume VI Part I* (192–210). Debrecen: Institute of English and American Studies University of Debrecen.

Howarth, P. A. (1996). *Phraseology in English Academic Writing* (Vol. 75). Tübingen: Max Niemeyer Verlag.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunyadi, L. (1999). Linguistic Analysis of Large Corpora: Approaches to Computational Linguistics in Hungary. *Literary and Linguistic Computing, 14* (1), 77–88.

Jackendoff, R. S. (1977). **X syntax: A study of phrase structure*. Cambridge, Mass.: The MIT Press.
Jamsa, K. (1997). *Így fejleszd a PC-det* (Váradi Z., Ford.). Budapest: Kossuth.
Johansson, S., Leech, G., & Goodluck, H. (1978). Manual information to accompany the Lancaster–Oslo/Bergen Corpus of British English, for use with digital computers [Online]. Elérhető: http://www.hd.uib.no/lob-www.htm
Johns, T. (1994). Contexts (Version 0.74).
Johns, T. (1997). Contexts: The background, development and trialling of a concordance-based CALL program. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Szerk.), *Teaching and language corpora* (100–115). London & New York: Longman.
Kam-mei, J. L., Chang, S., & James, G. (2003). A Glossary of Essential Academic Vocabulary. Letöltve: 2005. március 4. http://kdictionaries.com/newsletter/kdn11-09.html
Kenesei, I. (Szerk.). (2004). *A nyelv és a nyelvek* (5. javított, bővített kiadás). Budapest: Akadémiai Kiadó.
Kennedy, G. (1998). *An introduction to corpus linguistics*. London, England: Longman.
Kiefer, F. (1964a). A halmazelmélet egy nyelvészeti alkalmazásáról. In *Általános Nyelvészeti Tanulmányok I* (187–200). Budapest: Akadémiai Kiadó.
Kiefer, F. (1964b). Halmazelméleti és matematika-logikai modellek a nyelvben. In *Általános Nyelvészeti Tanulmányok I* (89–115). Budapest: Akadémiai Kiadó.
Kiefer, F. (1968). *Mathematical Linguistics in Eastern Europe*. New York: American Elsevier Publishing Co.
Kiefer, F. (Szerk.). (2003). *A magyar nyelv kézikönyve*. Budapest: Akadémiai Kiadó.
Kiss, Gábor (2004). A *piros, vörös* és más színnevek használata a Magyar Nemzeti Szövegtár alapján. In Gecső, T. (Szerk.), *Variabilitás és nyelvhasználat* (160–165). Budapest: TINTA Könyvkiadó.
Kjellmer, G. (1991). A mint of phrases. In K. Aijmer & B. Altenberg (Szerk.), *English corpus linguistics: Studies in honour of Jan Svartvik* (111–127). London: Longman.
Kjellmer, G. (1994). *A dictionary of English collocations: based on the Brown corpus*. Oxford: Clarendon Press, 1994
Knowles, G., Williams, B., & Taylor, L. (Szerk.). (1996). *A corpus of formal British English speech: The Lancaster/IBM Spoken English Corpus*. London & New York: Longman.
Kostić, Đ. (1965a). *Functions and Meanings of Cases in Serbo-Croatian*. Belgrád: Institute for Experimental Phonetics and Speech Pathology.
Kostić, Đ. (1965b). *Probability of Grammatical Forms in Serbo-Croatian*. Belgrád: Institute for Experimental Phonetics and Speech Pathology.
Kostić, Đ. (1965c). *Probability of Phonemic Co-occurrences of Serbo-Croatian Phonemes*. Belgrád: Institute for Experimental Phonetics and Speech Pathology.
Krishnamurthy, R. (1997a). 'Computers and Texts' [Course]. Debrecen, Hungary.
Krishnamurthy, R. (Ed.). (1997b). *Change and continuity at COBUILD (1986–96)* (Vol. 1). Eger, Hungary: Eszterházy Károly College.
Kugler, N., & Tolcsvai Nagy, G. (Szerk.). (2000). *Nyelvi fogalmak kisszótára*. Budapest: Korona.
Lager, T. (1995). *A logical approach to computational corpus linguistics*. [Göteborg]: Department of Linguistics Göteborg University.
Lea, D. (Szerk.). (2002). *Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press.
Lee, D. (2000). *Navigating through the BNC jungle using 'genre'*. Paper presented at Teaching and Language Corpora, Graz.
Leech, G. (1998). *English grammar in conversation.* Paper presented at Language Learning and Computers, Chemnitz University of Technology.
Leech, G., & Eyes, E. (1997). Syntactic annotation: treebanks. In R. Garside, Leech, G., & McEnery, T. (Szerk.), *Corpus Annotation: Linguistic Information from Computer Text Corpora* (34–52). London & New York: Longman.
Leech, G., Myers, G., & Thomas, J. (Szerk.). (1995). *Spoken English on computer: Transcription, mark-up and application* London: Longman.

Leitner, G. (1992a). International Corpus of English: Corpus design – problems and suggested solutions. In G. Leitner (Szerk.), *New directions in English language corpora: Methodology, results, software developments* (33–64). Berlin–New York: Mouton de Gruyter.

Leitner, G. (1992b). *New directions in English language corpora: methodology, results, software developments*. Berlin, New York, Mouton de Gruyter.

Levin, B., & Rappaport Havov, M. (1995). *Unaccusativity: At the syntax-lexical semantics interface*. Cambridge, MA: The MIT Press.

Linguistics (2005). *Encyclopaedia Britannica*. Letöltve: 2005. március 15. http://www.britannica.com/eb./article-3513/

Longman dictionary of contemporary English. (Új kiadás) (2003). Harlow, UK: Pearson Education Limited.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis & E. Tognini – Bonelli (szerk.), *Text and technology*: In honor of John Sinclair. Amsterdam: John Benjamins.

Matsuda, H., Tsuboi, Y., & Matsumoto, Y. (2001). VisualMorphs (Version 1.0). Nara: Nara Institute of Science and Technology.

Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Takaoka, K., et al. (1999). Japanese Morphological Analysis System ChaSen (Version 2.0). Nara: Nara Institute of Science and Technology, Japan.

McArthur, T. (Szerk.). (1992). *The Oxford companion to the English language*. Oxford: Oxford University Press.

McCarthy, M. (2004). *Touchstone: From corpus to course book*. Cambridge: Cambridge University Press.

McCarthy, M., McCarten, J., & Sandiford, H. (2005). *Touchstone: student's book 1*. Cambridge: Cambridge University Press.

McCartney, S. (1999). *ENIAC, the triumphs and tragedies of the world's first computer*. New York: Walker.

McEnery, T. (1992). *Computational linguistics: A handbook & toolbox for natural language processing*. Wilmslow, U.K.: Sigma Press.

McEnery, T., & Wilson, A. (1996). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Meijs, W. (1984). "You can do so if you want to" – Some elliptic structures in Brown and LOB and their syntactic description. In J. Aarts & W. Meijs (Szerk.), *Corpus linguistics: Recent developments in the use of computer corpora in English language research* (141–162). Amsterdam: Rodopi.

Micro-OCP. (1988). Oxford: Oxford University Press.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). Introduction to WordNet: An on-line lexical database. *Journal of Lexicography, 3*, 235–244.

Miller, G. A., & Fellbaum, C. (1992). Semantic networks of English. In B. C. Levin & S. Pinker (Szerk.), *Lexical & conceptual semantics*. Cambridge, MA: Blackwell.

Milroy, L. (1985). What a performance! Some problems with the competence-performance distinction. *Australian Journal of Linguistics, 5* (1–17).

MTA Nyelvtudományi Intézet. (1998–2002) Magyar Nemzeti Szövegtár. Letöltve: 2003. január 20. http://corpus.nytud.hu/mnsz/bevezeto_hun.html

Nadar, J. (1998). *Prentice Hall's illustrated dictionary of computing*. New York: Prentice Hall.

Nelson, G. (1995). The International Corpus of English: Mark-up for spoken language. In G. Leech, G. Myers & J. Thomas (Szerk.), *Spoken English on computer: Transcription, mark-up and application* (220–223) London: Longman.

Newmeyer, F. J. (1990). Competence vs. performance: Theoretical vs. applied; the development and interplay of two dichotomies in modern linguistics. *Historiographia Linguistica, 17*, 167–181.

Oakes, M. (1998). *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Ooi Vincent, B. Y. (1998). *Computer corpus lexicography*. Edinburgh: Edinburgh University Press.

Oravecz, C., & Dienes, P. (2002). Large scale morphosyntactic annotation of the Hungarian National Corpus. In *Studies in Linguistics Volume VI Part II: A supplement to the Hungarian Journal of English and American Studies* (277– 298). Debrecen: Institute of English and American Studies University of Debrecen.

Pajzs, J. (1990). Creating a historical dictionary of Hungarian with the aid of computer. In T. Magay & J. Zigány (Szerk.), *BUDALEX '88 Proceedings* (559–563). Budapest: Akadémai Kiadó.

Pajzs, J. (1991). The use of a lemmatized corpus for compiling the dictionary of Hungarian. In *Using corpora: Proceedings of the 7th Annual Conference of the OUP & Centre for the New OED and Text Research* (129–136): University of Waterloo Centre for the New OED.

Pajzs, J. (1994). Project report on the historical dictionary of Hungarian. In J. Pajzs, F. Kiefer & G. Kiss (Szerk.), *Papers in computational lexicography COMPLEX '94* (205–213). Budapest: Linguistics Institute, Hungarian Academy of Sciences.

Pajzs, J., Tihanyi, L., & Villó, I. (1992). Compiling dictionaries with grammar defined databases. In F. Kiefer, G. Kiss & J. Pajzs (Szerk.), *Papers in computational lexicography COMPLEX '92* (259–274). Budapest: Linguistics Institute, Hungarian Academy of Sciences.

Palmer, H. (1924). *A grammar of spoken English*. Cambridge: Heffer.

Papp, F. (Szerk.). (1966). *Mathematical linguistics in the Soviet Union*. The Hague: Mouton.

Papp, F. (Szerk.). (1969). *A magyar nyelv szóvégmutató szótára*. Budapest: Akadémiai kiadó.

Papp, F., Kulagina, O. S., & Melčuk, I. A. (1964). *Matematikai nyelvészet és gépi fordítás a Szovjetunióban*. Budapest.

Payne, J. (1995). The COBUILD spoken corpus: Transcription conventions. In G. Leech, G. Myers & J. Thomas (Szerk.), *Spoken English on computer: Transcription, mark-up and application* (203–207) London: Longman.

Payne, J. A., Sinclair, J., & Cobuild. (1995). *Collins Cobuild English guides 8: Spelling*. London: HarperCollins.

Peppé, S. (1995). The Survey of English Usage and the London-Lund Corpus: Computerizing manual prosodic transcription. In G. Leech, G. Myers & J. Thomas (Szerk.), *Spoken English on computer: Transcription, mark-up and application* (187–202). London: Longman.

Perkins, M. (1995). Corpora of disordered spoken language. In G. Leech, G. Myers & J. Thomas (Szerk.), *Spoken English on computer: Transcription, mark-up and application* (128–134) London: Longman.

Pfaffenberger, B. (1993). *QUE's computer user's dictionary* (4th ed.) Carmel, IN: QUE Corporation.

Piao, S. S. (2002). Multi-Lingual Corpus Toolkit (MLCT).

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.

Pinker, S. (2000). *Words and rules: The ingredients of language*. New York: Basic Boots.

Prószéky, G. (1989). *Számítógépes nyelvészet: Természetes nyelvek használata számítógépes rendszerekben*. Budapest: Számítástechnika-alkalmazási Vállalat.

Prószéky, G., & Kis, B. (1999). *Számítógéppel emberi nyelven: Intelligens szövegkezelés számítógéppel*. Bicske: Szak Kiadó.

Prószéky, G., & Tihanyi, L. (1992). A fast morphological analyzer for lemmatizing corpora of agglutinative languages. In F. Kiefer, G. Kiss & J. Pajzs (Szerk.), *Papers in computational lexicography COMPLEX '92* (275–278). Budapest: Linguistics Institute, Hungarian Academy of Sciences.

Prószéky, G., & Tihanyi, L. (1993). Humor: High-speed unification morphology and its applications for agglutinative languages. *La tribune des industries de la langue, 10*, 28–29.

Pustejovsky, J. (1995). *The generative lexicon.* Cambridge. Massachusetts: The MIT Press.

Pusztai, F. (Szerk.). (2003). *Magyar Értelmező Kéziszótár* (2. kiadás). Budapest: Akadémiai Kiadó.

The Random House Websters unabridged dictionary. (1997). (2. kiadás). New York: Random House.

Reed, A. (2003). Simple Concordance Program (SCP) (Version 4.0) [Win].

Renouf, A. (1987). Corpus development. In J. Sinclair (Szerk.), *Looking up* (1–40). London and Glasgow: Collins ELT.

Roach, P., & Arnfield, S. (1995). Linking prosodic transcription to the time dimension. In G. Leech, G. Myers & J. Thomas (Szerk.), *Spoken English on computer: Transcription, mark-up and application* (149–160) London: Longman.

Scott, M. (1996). Wordsmith Tools (Version 1). Oxford: Oxford University Press.

Scott, M. (1999). Wordsmith Tools (Version 3). Oxford: Oxford University Press.

Scott, M. (2004). Wordsmith Tools (Version 4). Oxford: Oxford University Press.

Scott, M., & Johns, T. (1993). MicroConcord. Oxford: Oxford University Press.

Sebba, M. (1995). Code switching: A problem for transcription and text encoding. In G. Leech, G. Myers & J. Thomas (Szerk.), *Spoken English on computer: Transcription, mark-up and application* (144–148). London: Longman.

Shastri, S. V. (1988). The Kolhapur Corpus of Indian English and work done on its basis so far. *ICAME Journal, 12*, 15–26.

Siki, Z. (1995). Adatbázis kezelés és szervezés. Letöltve: 2004. április 24. http://www.agt.bme.hu/szakm/adatb/adatb.htm.

Sinclair, J. (1991). *Corpus, concordance, collocation* (1st ed.). Oxford: Oxford University Press.

Sinclair, J. (1995). From theory to practice. In G. Leech, G. Myers & J. Thomas (Szerk.), *Spoken English on computer: Transcription, mark-up and application* (99–109). London: Longman.

Sinclair, J. (1996a). The empty lexicon. *International Journal of Corpus Linguistics, 1* (1), 99–119.

Sinclair, J. (Szerk.). (1987a). *Collins COBUILD English language dictionary*. London: Collins.

Sinclair, J. (Szerk.). (1987b). *Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary*. London and Glasgow: Collins ELT.

Sinclair, J. (Szerk.). (1990). *Collins COBUILD English grammar*. London: HarperCollins.

Sinclair, J. (Szerk.). (1996b). *Collins COBUILD Grammar Patterns 1: Verbs*. London: HarperCollins Publishers.

Sinclair, J. (Szerk.). (1998). *Collins COBUILD Grammar Patterns 2: Nouns and adjectives*. London: HarperCollins Publishers.

Sinclair, J., & Cobuild. (1991a). *Collins Cobuild English guides 1: Prepositions*. London: HarperCollins.

Sinclair, J., & Cobuild. (1991b). *Collins Cobuild English guides 2: Word formation*. London: HarperCollins.

Sinclair, J., & Cobuild. (1995). *Collins Cobuild English guides 6: Homophones*. London: HarperCollins.

Sinclair, J., Jones, & Daley. (1969). *English lexical studies* (No. OSTI Report).

Sinclair, J. M. (1993). Text corpora: Lexicographers' needs. *Zeitschrift fur Anglistik und Amerikanistik, 1*, 6–14.

Sinclair, J. M. (2001). Preface. In M. Ghadessy, A. Henry & R. L. Roseberry (Szerk.), *Small corpus studies and ELT: Theory and practice*. Amsterdam: John Benjamins.

Sinclair, J. M. (2004). *How to use corpora in language teaching* (Vol. 12). Amsterdam: John Benjamins Publishing Company.

Sperberg-McQueen, C. M. (1994). The Text Encoding Initiative. In A. Zampolli, M. Palmer & N. Calzolari (Szerk.), *Current issues in computational linguistics: In honour of Don Walker* (409–428). Dordrecht: Kluwer Academic Publishers.

Sperberg-McQueen, C. M., & Burnard, L. (1994). *Guidelines for electronic text encoding and interchange*. Chicago, Oxford: Text Encoding Initiative.

Stubbs, M. (1996). *Text and corpus analysis: computer assisted studies of language and institutions*. Oxford: Blackwell Publishers.

Szépe, G. (2001). Bevezető szavak. In M. Barkó-Nagy, Z. Bánréti & K. É. Kiss (Szerk.), *Újabb tanulmányok a strukturális magyar nyelvtan és a nyelvtörténet köréből: Kiefer Ferenc tiszteletére barátai és tanítványai* (9–16). Budapest: Osiris Kiadó.

Tankó, G. (2004). Composition: The use of adverbial connectors in Hungarian university students' argumentative essays. In J. M. Sinclair (Szerk.), *How to Use Corpora in Language Teaching* (157–181). Amsterdam: John Benjamins Publishig Company.

Taylor, D. S. (1988). The meaning and use of the term 'competence' in linguistics and applied linguistics. *Applied Linguistics, 9*, 148–168.

Thompson, G. (1995). *Reporting*. London: HarperCollins.

Thompson, G., & Cobuild. (1994). *Collins Cobuild English guides 5: Reporting*. London: HarperCollins.

Thompson, H. S., Anderson, A. H., & Bader, M. (1995). Publishing a spoken and written corpus on CD-ROM: The HCRC Map Task experience. In G. Leech, G. Myers & J. Thomas (Szerk.), *Spoken English on computer: Transcription, mark-up and application* (168–180) London: Longman.

Thompson, P., Sealey, A., & Scott, M. (2004). Kids, corpora and concordancing, *The sixth Teaching and Language Corpora conference*. Granada, Spain.

Tottie, G., Eeg-Olofsson, M., & Thavenius, C. (1984). Tagging negative sentences in LOB and LLC. In J. Aarts & W. Meijs (Szerk.), *Corpus linguistics: Recent developments in the use of computer corpora in English language research* (173–184). Amsterdam: Rodopi.

Tribble, C., & Jones, G. (1990). *Concordances in the classroom*. Harlow: Longman.

Valera, S., & Rizo-Roderiguez, A. (1998). A LOB-Corpus-based semantic profile of the adjective in English supplementive clauses. *International Journal of Corpus Linguistics, 3* (2), 251–278.

Váradi, Tamás & Oravecz, Cs. (1999). Morpho-syntactic ambiguity and tagset design for Hungarian. In *Proceedings of the Workshop on Linguistically Interpreted Corpora EACL '99* (8–13). Bergen.

Váradi, Tamás (2000). Lexical and Translation Equivalence in Parallel corpora. In *Proceedings of the Second International Conference on Language Resources and Evaluation* ELRA, ELRA I. (539–543). Athens.

Váradi, Tamás (2000). Fishing for Translation Equivalents Using Grammatical Anchors. *International Journal of Corpus Linguistics,* 5/1, (1–16).

Váradi, Tamás & Kiss, Gábor (2001). Equivalence and Non-equivalence in parallel corpora. *International Journal of Corpus Linguistics,* 6 (special issue) (166–177).

Váradi, Tamás (2001). The Linguistic Relevance of Corpus Linguistics. In McEnery, T. et al. (eds.) *Proceedings of Corpus Linguistics,* UCREL, Lancaster.

Váradi, Tamás (2002a). The Hungarian National Corpus, in *Proceedings of the Third International Conference on Language Resources and Evaluation* ELRA, ELRA: Las Palmas.

Váradi, Tamás (2002b). A nyelvhasználat empirikus vizsgálatáról. In Andor J., Szűcs T. & Terts I. (Szerk.) *Színes eszmék nem alszanak*. (1285–1291). Pécs: Lingua Franca csoport.

Váradi, T. (2003). *The shallow parsing of Hungarian business news.* Paper presented at the Proceedings of Workshop on Shallow Processing of Large Corpora (SProLaC03), Lancaster.

Warren, L. (1992). Learning from the learners' corpus. *Modern English Teacher, 1*, 9–11.

Watson, J., Batten, J., Willis, D., Cobuild, & University of Birmingham. (1991). *Collins COBUILD student's grammar*. London: HarperCollins.

Watt, R. J. C. (2004). Concordance (Version 3.2).

Wehmeier, S. (Szerk.). (2000). *Oxford advanced learner's dictionary of current English* (6. kiadás). Oxford: Oxford University Press.

Wehmeier, S. (Szerk.) (2005). Oxford: Oxford University Press. (7. kiadás)

Weitzenbaum, J. (1976). Eliza.

West, M. P., & Endicott, J. G. (1935). *The New Method English Dictionary*. London: Longmans, Green.

Wichmann, A., Fligelstone, S., McEnery, T., & Knowles, G. (Szerk.). (1997). *Teaching and language corpora*. London: Longman.

Wikberg, K. (1992). Discourse category and text type classification: Procedural discourse in the Brown and the LOB corpora. In G. Leitner (Szerk.), *New directions in English language corpora: Methodology, results, software developments* (247–262). Berlin–New York: Mouton de Gruyter.

Willis, D. (1990). *The lexical syllabus: A new approach to language teaching*. Glasgow: HarperCollins Publishers.

Willis, D., & Wright, J. (1995). *Collins COBUILD basic grammar*. London: HarperCollins.

Willis, J. R., & Willis, J. D. (1988). *Collins Cobuild English Course.* London: Collins.

TÁRGYMUTATÓ

adatbázis 11, 18–22, 37, 50, 86, 87, 91, 92, 98, 164, 167, 171, 182
adatvezérelt nyelvtanulás 167
ágbank 167
ágrajz 9, 42, 68, 167
általános korpusz 27, 33, 167, 169
annotáció 9, 17, 37, 38, 40, 42–44, 46, 62, 67, 73, 100–102, 167, 168
annotált korpusz 42, 100, 102, 129, 167
AntConc 9, 10, 103, 113, 127–129, 176
átírás 37, 38, 62, 90–92, 167, 168
automatikus szintaktikai elemzés 168
automatikus szintaktikai elemző program 168
beszédfelismerés 22, 54, 60, 168
címke 43, 168
címkézés 18, 38
címkéző program 66, 101, 160, 168
COCOA utalások 168
ConcApp 10, 113, 114, 126, 129, 178
Contexts 10, 15, 158–160, 176, 179
csontváz elemzés (szintaktikai) 168
dinamikus korpusz 32, 168
előfordulás 9, 20, 23, 29, 30, 65, 68, 83, 84, 102, 104, 105, 108–111, 118, 119, 123, 124, 127, 128, 144
fordítási ekvivalencia / megfelelés 168
fordítási korpusz 34, 169
Hapax legomenon / legomena 29, 169
homográf 101, 169
idióma elv 169, 172
jelentés egyértelműsítés 169
kiegyensúlyozott korpusz 169
klón 22, 34, 49, 74, 75, 169
kolligáció 169, 170
kollokáció 72, 108–110, 118, 119, 134, 138, 147, 151, 152, 158, 169, 171, 173, 174
konkordancia 9, 10, 18, 52, 82–84, 89, 90, 103, 104, 112, 119, 120, 123, 127–129, 133, 140, 144–147, 149, 151–161, 167, 170, 174
konkordanciaprogram 16, 102, 104, 109, 110, 112–114, 129, 140, 170
kontexus 9, 35, 52, 63, 83, 102, 112
korpusz 9, 11, 17, 18–20, 22–29, 30–39, 42, 43, 45–47, 50, 52, 56, 58, 60–82, 86–95, 97–102, 112, 129, 130, 132–135, 138, 140, 142–145, 157, 162–174, 178, 184
korpusz alapú 27, 28, 58, 129, 133, 138, 142–144, 170
korpusz informált 170
korpusznyelvészet 13, 15–19, 44, 47, 50, 51, 53, 59, 64, 98, 130, 132, 133, 141, 144, 162, 170
korpuszvezérelt 140, 170, 171
KWAL formátum 170
KWIC formátum 170
lemma 30, 154, 170, 171
lemmatizálás 30, 52, 170
lemmatizáló program 171
lexika 134, 171
lexikon 9, 30, 52, 171
mesterséges intelligencia 53, 54, 59, 171, 174
MI együttható 171
mintavétel 17, 23, 32, 34, 46, 74, 171
MLCT 10, 113–116, 129, 130, 182
monitor korpusz 32, 46, 171
n-gram 117, 128, 171
nyelvtanulói korpusz 34, 70, 165, 167, 171
összehasonlítható korpusz 34, 171
párhuzamos korpusz 34, 164, 169, 172–174
pedagógiai korpusz 35, 172
példány 28, 172
probabilisztikus módszerek 60, 172
regexp (reguláris kifejezés) 114, 119, 172
reprezentatív 18, 19, 23, 25, 27, 46, 67, 80, 172
SGML szabványos, általános leíró nyelv 172
Simple Concordance Program 113, 120, 181
statikus korpusz 32, 172
statisztikai jelentősség 172
szabad választás elve 169, 172
számítógépes nyelvészet 15, 47, 53–57, 59, 89, 172, 181
szemantikai prozódia 154, 155, 173
szintaktikai adatbázis 167
szóalak/típus 28, 86, 88, 106, 118, 124, 125, 168–171, 173
szókincstár 54, 173
szövegarchívum 19, 166, 173
szöveggyűjtemény 19, 20, 173
szövegkörnyezet 10, 29, 97, 102, 108, 109, 119, 152, 154, 170, 173

szövegszinkronizálás 173
szövegszó 9, 25, 28, 33, 35, 63, 80, 81, 86–90, 101, 123, 124, 172, 173
T-együttható 173
TEI (szövegkódolási ajánlás) 42, 88, 173
teljes szintaktikai elemzés 173
természetes nyelvfeldolgozás 58, 173
thesaurus 173
többnyelvű korpusz 34, 174
többváltozós statisztikai elemzés 174
történeti/diakronikus korpusz 35, 46, 77, 81–83, 112, 166, 174
tudásbázis 174
tudatosság javítása / tudatosságot javító 174
vizsgált csomópont / adat (nód) 174
WordCruncher 111, 176
Z-együttható 174

NÉVMUTATÓ

Aarts, B. 176
Aarts, J. 51, 176–178, 180, 182
Aijmer, K. 34, 176, 179
Altenberg, B. 34, 176, 179
Anderson, A. H. 38, 182
Andor, J. 12, 51, 176
Anthony, L. 9, 113, 176, 181
Arnfield, S. 38, 181
Aston, G. 132, 176
Atkins, S. 23, 176
Atwell, E. 64, 176
Bach, I. 57, 176
Bader, M. 38, 182
Bánréti Z. 182
Bakró-Nagy, M. 182
Barlow, M. 88, 89, 113, 176
Batten, J. 183
Bauer, L. 63, 74, 176
Beckwith, R. 180
Benson, E. 110, 147, 176
Benson, M. 110, 147, 176
Berry, R. 140, 176
Biber, D. 30, 64, 78, 142, 143, 176
Bolla K. 56
Bottyán, G. 81, 176
Burnard, L. 43, 67, 132, 177, 182
Bussmann, H. 53, 177
Calzolari, N. 182
Capel, A. 140, 177
Carpenter, E. 140, 177
Chalker, S. 140, 177
Chang, L. 179
Cheepen, C. 38, 177
Cheng, W. 27, 177
Chomsky, N. 51, 59, 176, 177
Clear, J. 19, 176, 177
Cobuild 140, 176–179, 181–183
Codd, E. F. 20, 177
Conrad, S. 176
Cook, G. 38, 177
Cowie, A. P. 134, 177
Crowdy, S. 38, 177
Crystal, D. 19, 109, 142, 167, 177
Daley 182
Deignan, A. 140, 177
Dienes, P. 44, 180
Dodd, B. 10, 89, 155, 177
É. Kiss K. 182
Edwards, J. A. 38, 177
Eeg-Olofsson, M. 64, 182
Endicott, J. G. 134, 183
Eyes, E. 43, 179
Fang, A. C. 72, 177
Fellbaum, C. 21, 52, 177, 180
Fillmore, C. J. 51, 52, 177
Finegan, E. 30, 176
Firth, J. R. 52, 110, 177, 178
Fligelstone, S. 179, 183
Francis, G. 180
Francis, W. N. 19, 24, 63, 175, 178
Fromkin, V. 52, 178
Garside, R. 38, 64, 176, 178, 179
Gatenby, A. V. 178
Gecső, T. 179
Ghadessy, M. 132, 178, 182
Goldstine, H. H. 47, 48l, 178
Goodale, M. 140, 178
Goodluck, H. 179
Granger, S. 71, 132, 178
Greaves, C. 113, 178
Greenbaum, S. 63, 67, 74, 178
Gross, D. 180
Habert, B. 91, 178
Halliday, M. A. K. 52, 178
Hann 177
Haslerud, V. 38, 178
Henry, A. 178, 182
Hirano, Y. 180
Hollósy B. 12, 175, 178
Hornby, A. S. 134, 178
Horváth J. 34, 65, 72, 73, 178
Howarth, P. A. 151, 178
Hung, J. 178
Hunston, S. 12, 17, 178
Hunyadi L. 56, 178
Ilson, R. 176
Jackendoff, R. S. 52, 179
James, G. 79, 179
Jamsa, K. 48, 179
Johansson, S. 176, 178, 179

Johns, T. 10, 15, 111, 144, 146, 155, 158, 167, 179, 181
Jones, G. 111, 182
Kam-mei, J. 72, 179
Kenesei I. 109, 179
Kennedy, G. 61, 64, 68, 78, 179
Kiefer F. 56, 59, 109, 179, 181, 182
Kis B. 57, 86, 181
Kiss G. 12, 56, 81, 179, 181
Kiss-Gulyás J. 178
Kitauchi, A. 180
Kjellmer, G. 30, 110, 179
Knowles, G. 37, 179, 183
Kostić, Đ. 9, 60, 61, 95, 179
Krishnamurthy, R. 11, 12, 23, 28, 64, 65, 139, 179
Kučera, H. 24, 63, 178
Kugler N. 17, 109, 167, 179
Kulagina, O. S. 181
Lager, T. 11, 51, 52, 179
László Zs. 177
Lea, D. 138, 179
Leech Geoffrey, N. 9, 38, 43, 64, 79, 142, 176–182
Leitner, G. 63, 74, 177, 179, 180, 183
Levin, B. 52, 180
Louw, B. 154, 180
Magay T. 180
Matsuda, H. 45, 180
Matsumoto, Y. 44, 180
McArthur, T. 18, 180
McCarten, J. 180
McCarthy, M. 31, 135, 143, 180
McCartney, S. 47, 180
McEnery, T. 9, 29, 30, 54, 78, 119, 132, 160, 177, 179, 180, 183
Meijs, W. 17, 64, 176, 180, 182
Melčuk, I. A. 181
Miller, G. A. 21, 52, 180
Miller, K. 180
Milroy, L. 52, 180
Myers 38, 177, 178–182
Nadar, J. 50, 180
Nazarenko, A. 178
Nelson, G. 38, 180
Newmeyer, F. J. 52, 180
Nikléczy P. 56
Oakes, M. 119, 180
Olaszy G. 56
Ooi Vincent, B. Y. 119, 180
Oostdijk 176–178
Oravecz C. 44, 81, 180, 183
Ostler, N. 176
Pajzs J. 56, 59, 181
Palmer, H. 110, 134, 142, 181, 182
Papp F. 56, 59, 181
Payne, J. 38, 140, 181
Peppé, S. 38, 181
Perkins, M. 38, 181
Petch-Tyson, S. 178
Pfaffenberger, B. 54, 181
Piao, S. S. 113, 181
Pinker, S. 52, 180, 181
Prószéky G. 44, 57, 59, 181
Tihanyi L. 44, 57, 181
Pustejovsky, J. 52, 181
Pusztai F. 134, 181
Rappaport Havov, M. 52, 180
Rebrus P. 177
Reed, A. 113, 181
Renouf, A. 64, 181
Rizo-Roderiguez, A. 64, 183
Roach, P. 38, 181
Roseberry, R. L. 178, 182
Salem, A. 178
Sandiford, H. 180
Scott, M. 9, 106, 111, 112, 155, 158, 181, 182
Sealey, A. 158, 182
Sebba, M. 38, 181
Shastri, S. V. 63, 74, 182
Siki Z. 20, 182
Sinclair, J. 10, 12, 19, 22, 29, 30, 32, 38, 50, 52, 64, 73, 75, 132, 134, 140, 141, 169, 172, 175, 180–182
Sperberg-McQueen 43, 182
Stenström, A.-B. 38, 178
Stubbs, M. 52, 182
Svartvik, J. 177–179
Szemere P. 177
Szentgyörgyi S. 177
Szentgyörgyi-Kiss K. 177
Szépe Gy. 56, 182
Szűcs T. 177, 183
Takaoka, K. 180
Tankó Gy. 73, 182
Taylor, D. S. 52, 182
Taylor, L. 179
Terts, I. 183
Thavenius, C. 64, 182
Thomas, J. 38, 177–182
Thompson, G. 140, 182
Thompson, P. 38, 158, 182
Tihanyi L. 44, 57, 181
Tolcsvai Nagy G. 17, 109, 167, 179
Tottie, G. 64, 182
Tribble, C. 111, 182

Tsuboi, Y. 180
Valera, S. 64, 183
Váradi T. 42, 56, 81, 183, 183
Villő I. 181
Vinkler Zs. 177
Wakefield, H. 178
Warren, L. 70, 183
Warren, M. 27, 177
Watson, J. 140, 183
Watt, R. J. C. 113, 183
Wehmeier, S. 134, 183
Weitzenbaum, J. 54, 183
West, M. P. 134, 183
Wichmann, A. 179, 183
Wikberg, K. 64, 183
Williams, B. 179
Willis, D. 35, 65, 140, 143, 172, 183
Willis, J. R. 12, 139, 143, 183
Wilson, A. 9, 29, 30, 78, 119, 180
Wright, J. 140, 183
Yamashita, T. 180
Zampolli, M. 182
Zigány J. 180
Zólyomi G. 177

KORPUSZOK MUTATÓJA

American National Corpus 22, 164
American Printing House for the Blind (APHB) 9, 29, 30, 164
Australian Corpus of English (ACE) 34, 63, 74, 164
Bank of English 11, 33, 64, 65, 78, 164
Base de données textuelles, ChroQué 92, 164
Bergen Corpus of London Teenager Language (COLT) 38, 164, 178
BESEDA 97, 164
British National Corpus (BNC) 9, 31, 33, 38–40, 66, 67, 76, 79, 98, 164, 165, 177–179
Brown University Standard Corpus of Present-Day American English 50, 63, 164
Brooklyn–Geneva–Amsterdam–Helsinki Parsed Corpus of Old English 77, 164
Cambridge International Corpus (CIC) 76, 135, 164
Cambridge and Nottingham Corpus of Discourse in English (CANCODE) 23, 31, 33, 76, 164
CHILDES Database 87, 164
COBUILD 9, 28, 29, 31, 33, 38, 64, 65, 70, 75, 139, 140, 141, 164, 178, 179, 181, 182, 183, 184
Corpus of English-Canadian Writing 34, 63, 74, 80, 164
Corpus du Théâtre religieux français du Moyen Âge 93, 164
Corpus VALIFLOUI 92, 164
Cseh Nemzeti Korpusz 98, 164
Dialogstrukturenkorpus 90, 164
English–Swedish Parallel Corpus 34, 164, 176
Eötvös Loránd Tudományegyetem Korpusza 73, 165
EuroWordNet 165
FIDA 97, 165
FrameNet 52, 165
Francia Beszélt Nyelvi Korpusz 91, 165
Freiburg Corpus Freiburg–Brown (FROWN) 11, 63, 74, 75, 79, 165
Freiburg–LOB (FLOB) 11, 63, 74, 75, 165
Freiburger Korpus 90, 165
Hansard Corpus 78, 165
Hong Kong University of Science and Technology (HKUST) Corpus of Learner English 35, 72, 79, 165
Hong Kong Corpus of Conversational English (HKCCE) 27, 33, 165, 177
Horvát Nemzeti Korpusz 95, 165
Human Communication Research Centre's Map Task Corpus 38, 165
Hunglish Korpusz 11, 87, 165
International Computer Archive of Modern and Medieval English, ICAME 10, 35, 62, 63, 75, 78–80, 141, 165, 177, 182
International Corpus of English (ICE) 9–11, 25–27, 34, 38, 63, 67–69, 74, 79, 165, 178–180
International Corpus of Learner English (ICLE) 11, 34, 70, 71, 165, 178
Janus Pannonius Tudományegyetem Korpusza (JPU Corpus) 72, 165
Japán diákok angol nyelvű korpuszai 72, 165
Kanadai Farncia Korpusz 91, 165
Kolhapur Corpus of Indian English (KOL) 34, 63, 74, 79, 165, 182
Lancaster–IBM Spoken English Corpus 37, 165
Lancaster–Leeds Treebank 165
Lancaster–Oslo/Bergen Corpus (LOB) 23, 24, 30–34, 63, 64, 68, 74, 79, 111, 165, 176, 180, 182, 183
Linguistic Data Consortium, LDC 19, 98, 165
London–Lund Corpus of Spoken English (LLC) 31, 38, 62, 64, 165, 182
Longman Corpus Network (LCN) 70, 76, 165
Magyar dalszövegek 87, 165
Magyar Elektronikus Könyvtár 9, 36, 37, 81, 87, 114, 166
Magyar Irodalmi és Köznyelv Nagyszótárának Korpusza / Magyar Történeti Korpusz 9, 81–83, 112
Magyar Nemzeti Szövegtár (MNSZ) 10, 11, 27, 31, 33, 39, 80, 81, 99, 110, 133, 145, 149, 151, 154, 155, 157, 166, 180
Magyar Webkorpusz 87, 166
negr@ korpusz 89, 166
Német telefonbeszélgetések 166
Oxford Text Archive 19, 78, 166
PAROLE Francia Korpusz 91, 166
Parsed Corpus of Early English Correspondence 77, 166

PELCRA (Polish and English Language Corpora for Research and Applications) 98, 166
Penn–Helsinki Parsed Corpus of Early Modern English 77, 166
Penn–Helsinki Parsed Corpus of Middle English 35, 77, 166
Pfeffer–Korpus 90, 166
Survey of English Usage (SEU) 11, 38, 60–62, 68, 79, 166, 181
Szerb Nyelv Korpusza 60, 93, 166
Szeged Korpusz 11, 86, 166
Tiger Korpusz 89, 166
Tycho Brahe Parsed Corpus of Historical Portuguese 35, 166
Wellington Corpus of Written New Zealand English 34, 63, 74, 166
WordNet 9, 21, 22, 52, 166
World English Corpus 77, 166
York–Helsinki Parsed Corpus of Old English Poetry 77, 166
York–Toronto–Helsinki Parsed Corpus of Old English Prose 77, 166