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.

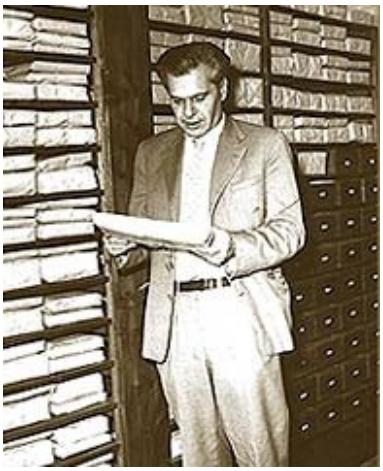
3.2. Az elektronikus korpuszok előfutárai

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.

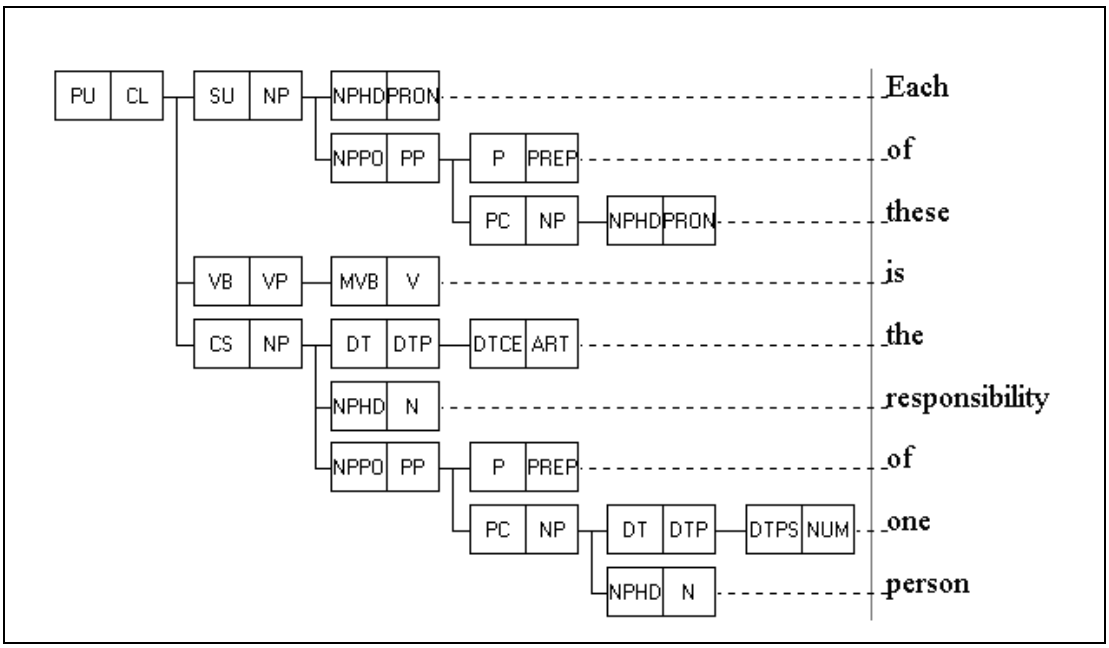
A további kritikák között említsük meg Lou Burnard előadását a 4. Nemzetközi Oktatás és Nyelvi Korpuszok Konferenciáról (Fourth International Conference on Teaching and Language Corpora – TALC), amelyben a BNC tervezéséről és bizonyos döntésekről szólt (L. Burnard, 2000). Már a cím is sokat sejtető: „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		szépprózа					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		számítás-technikai 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

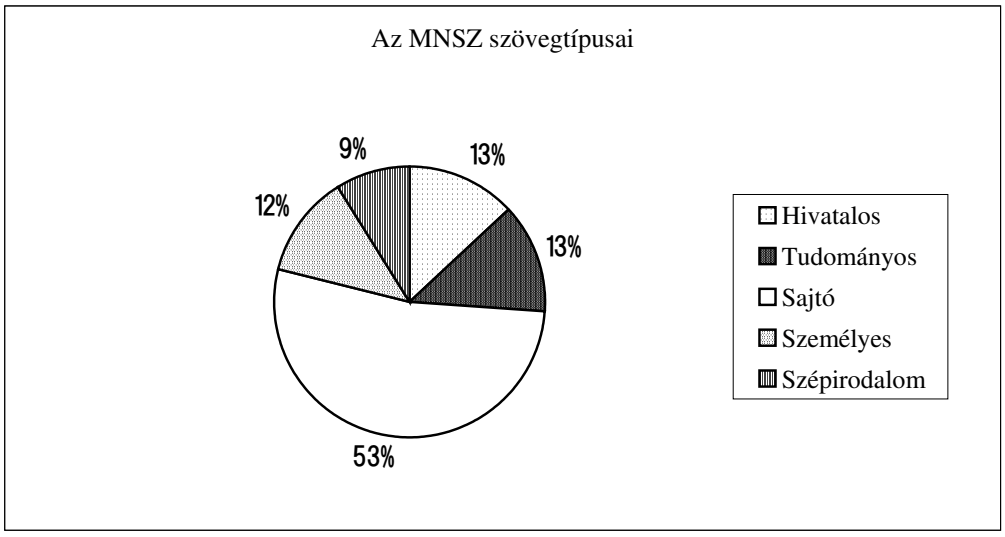
		Név	Nyelv	Tartalom	Méret (szószám)	Típus	Címkézett	Elemzett	Elérhetőség/ illetékes
Egynyelvű korpuszok	Írott	The Longman-Lancester Corpus	brit, amerikai, nyugat-afrikai angol		30 000 000				Della Summers
		The Macquaire Corpus	ausztráliai angol		1 000 000				Pam Peter
		The Susanne Corpus		Brown része	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:



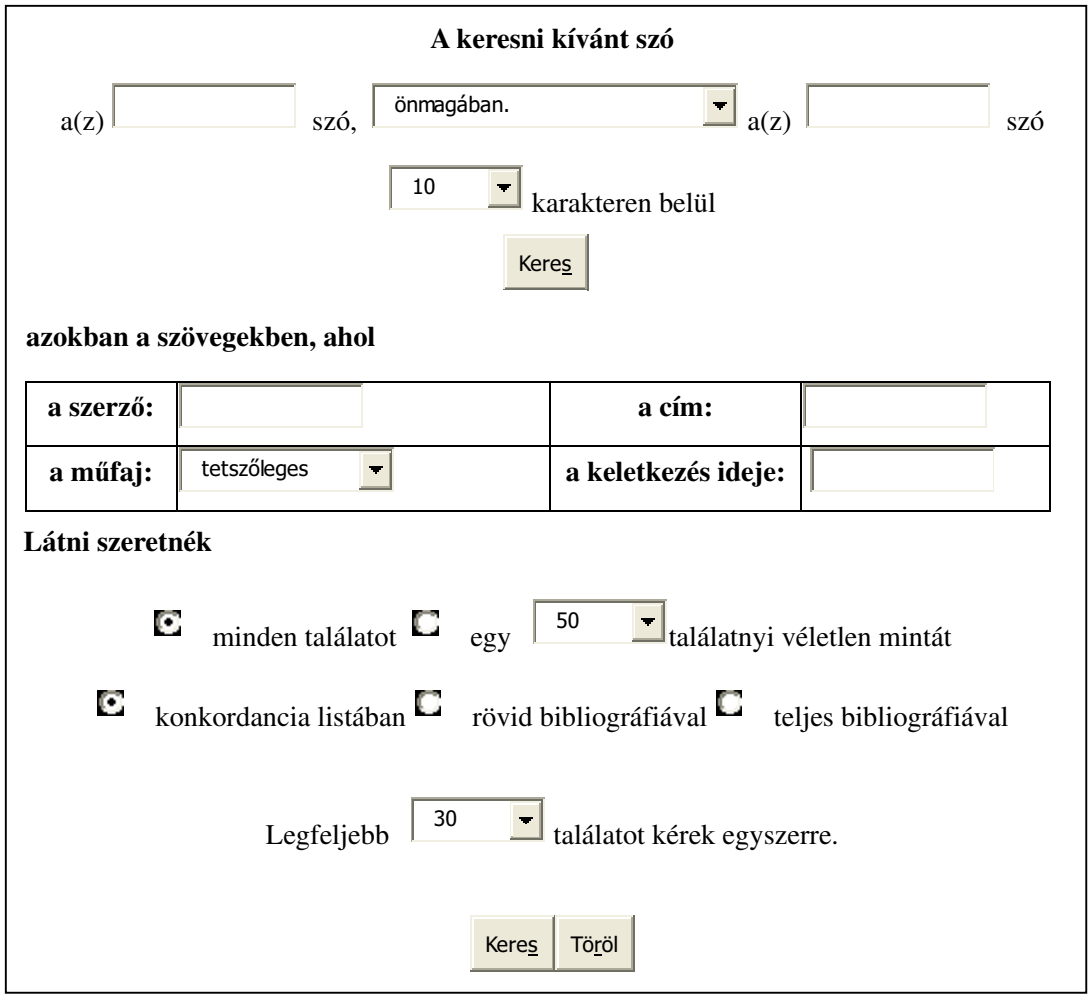
21. ábra: Az MNSZ szövegtípusainak szövegszó szerinti aránya

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 orcáját; /
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:

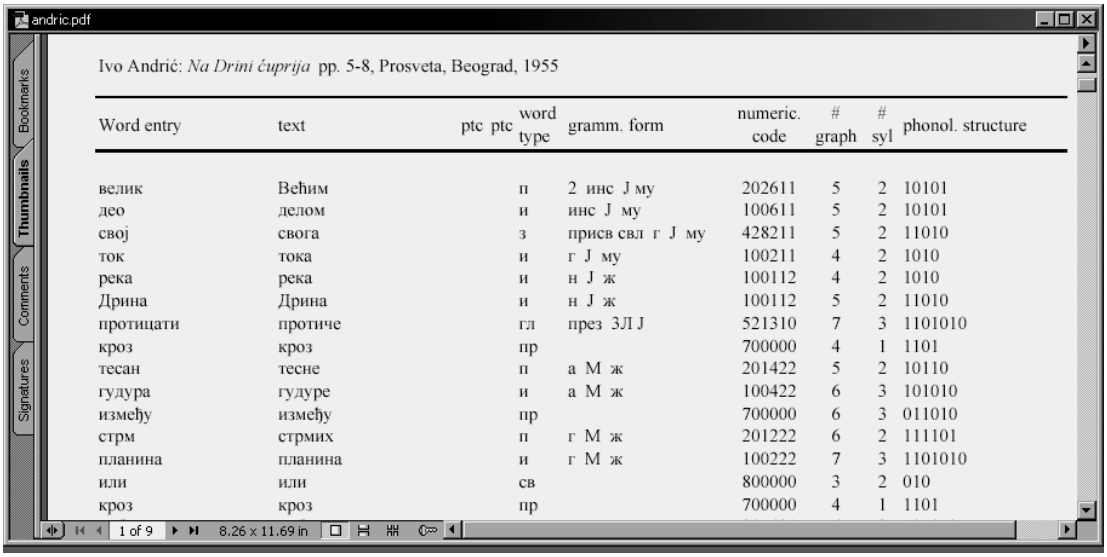
12^{th} – 18^{th} century: the list of books and items **CSL**

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.



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

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></td><td>SYN2000
100 millió</td><td>ORAL PMK
Prágai beszélt nyelv</td><td>DIAL
tervezett dialektális</td><td></td><td></td><td>DIAKORP</td><td>DIAL
tervezett dialektikus</td><td></td></tr>
<tr><td></td><td>PUBLIC
20 millió</td><td></td><td></td><td></td><td></td><td></td><td></td><td></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.