Bevezetés a korpuszok csodálatos világába

Nyelvi adatok feldolgozása – 2019/20 tavasz
4. óra

Simon Eszter

MTA Nyelvtudományi Intézet

Tartalom

Mi a korpusz?

Mi a korpusz? 1.

Kugler és Tolcsvai Nagy (2000)

„meghatározott szempontok alapján kiválasztott szövegmennyiség, amelyen a nyelvész vizsgálatát végzi”

A főbb kérdések

- mennyiség
- nyelvészeti vizsgálatokra alkalmas
- reprezentativitás, a kiválasztás szempontjai
- tárolás módja: elektronikus
- tartalom: szegmentálás, annotáció, metaadatok

Mi a korpusz? 2.

Sinclair (2005)

„a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research”

Korpusz vs. adatbázis

ID	betűhű	normalizált	tő	elemzés
1	Athyanak	Atyának	Atya	N:P.Dat_gen
2	ees	és	és	C
3	fywnak	Fiúnak	Fiú	N:P.Dat_gen
4	ees	és	és	C
5	zenth leleknek	Szentléleknek	Szentlélek	N:P.Dat_gen
6	newe-@@ben	nevében	név	N.PxS3.Ine

ID	szerző	cím	év	hely	kiadó
1	Nietzsche	Az Antikrisztus	1993	Budapest	Ictus
2	Nietzsche	Ecce homo	1993	Budapest	Göncöl
3	Nietzsche	Bálványok alkonya	2004	Budapest	Holnap

Korpusztipológia

Modalitás/forrás

- írott
- hangzó (audio) (paasonen_1315.eaf)
- video (http://szotar2.jelesely.hu/index.php?word=2199)
- multimodális (pl. gesztusfelismerés, prozódia, diskurzuselemzés)
- kézzel írott, nyomtatott, eleve elektronikusan keletkezett

Domén/genre/műfaj

- gazdasági rövidhírek
- termékleírások
- szoftverdokumentáció
- szépirodalom
- diákfogalmazások
- tudományos írások
- enciklopédia
- ...

Nyelv

- egynyelvű
- kétnyelvű
- többnyelvű

párhuzamos korpuszok (parallel corpora)

a forrásnyelvi szöveget (S) és annak célnyelvi fordítását (T) tartalmazzák, mondat- vagy bekezdésszinten párhuzamostíva $\rightarrow$ S és T pontos fordítása egymásnak

összevethető korpuszok (comparable corpora)

ha S és T nem pontos fordításai egymásnak, de a mintavétel módját tekintve megegyeznek, akkor beszélünk összevethető korpuszról (McEnery és Xiao, 2007)

Származási hely

US	Brown Corpus
UK	Lancester–Oslo/Bergen Corpus
India	Kolhapur Corpus of Indian English
Ausztrália	Australian Corpus of English
Új-Zéland	Wellington Corpus of Written New Zealand English
Kanada	Corpus of English-Canadian Writing

Származási idő

- szinkrón (Magyar Nemzeti Szövegtár)
- diakrón (Ómagyar Korpusz)

Gold standard korpusz

Követelmények:

- kihalt nyelvek esetében kimerítő, amúgy reprezentatív, de legalábbis kiegyensúlyozott
- nyelvi elemekre van bontva (token, mondat, bekezdés…)
- nyelvi annotáció van minden elemhez rendelve
- az annotáció vagy kézzel készül, vagy kézzel van ellenőrizve egy előre kidolgozott annotációs séma és útmutató alapján
- jellemzően előre meghatározott a mérete

Silver standard korpusz

- maga a korpusz vagy az annotáció automatikusan generált
- az annotáció megbízhatósága fontos szempont

hunNERwiki corpus

- 19 millió token
- a magyar Wikipédiából a DBPedia kategóriáit felhasználva
- 92,94%-os F-mérték
- http://hlt.sztaki.hu/resources/hunnerwiki.html

Mire jó a korpusz?

Mire jó a korpusz?

- szinkrón nyelvi jelenségek vizsgálatára
- longitudinális nyelvészeti vizsgálatokra
- nyelvtanulásra
- nyelvfeldolgozó eszközök tanítására és tesztelésére
- szótárépítésre
- ...

Főbb kérdések a korpuszépítésnél

Korpuszdizájn

Tisztázandó kérdések:

- kik és mire fogják használni a korpuszt
- a nyelvváltozat, amit le szeretnénk fedni
- a műfaj, amit reprezentálni szeretnénk
- a szükséges méret
- a korpusz jövőbeli elérhetősége, használhatósága → copyright kérdések és a szöveggyűjtés nehézségei

Mintavételezés, reprezentativitás

McEnery (2004)
„collected within the boundaries of a *sampling frame* designed to allow the exploration of certain linguistic feature (or set of features) via the data collected”

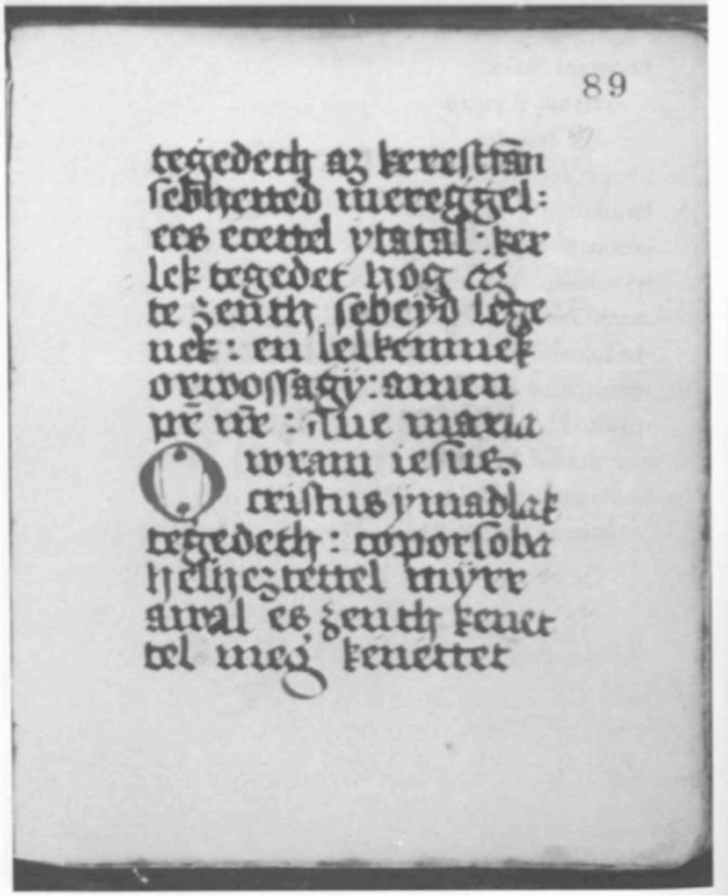
Hunston (2008)
„*representativeness* is the relationship between the corpus and the body of language it is used to represent”

Szöveggyűjtés

A szöveg forrása:

- elektronikus formátum
  - gép által olvasható, strukturált szöveges formátum → XML-parszolás
  - strukturálatlan szöveges formátum → strukturálttá alakítás
  - kép → szöveggé alakítás
- papíralapú formátum → elektronikussá alakítás

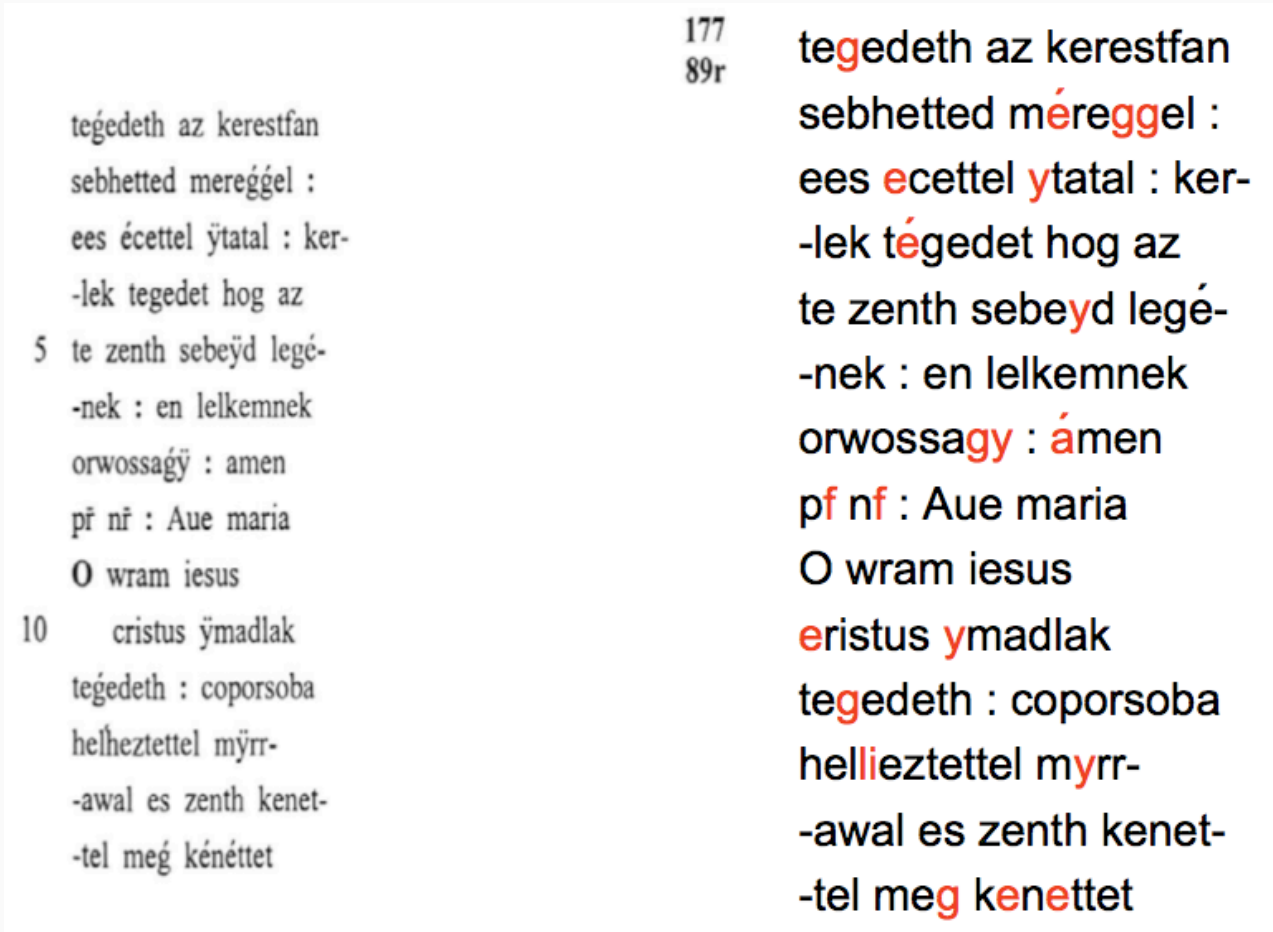
Kézzel írott kódexek, nyomtatott kiadások



177
89r

teǵedeth az kerestfan
sebhetted mereǵǵel :
ees écettel ÿtatal : ker-
-lek tegedet hog az
te zenth sebeÿd legé-
-nek : en lelkemnek
orwossagÿ : amen
pĭ nĭ : Aue maria
O wram iesus
cristus ÿmadlak
teǵedeth : coporsoba
helheztettel mÿrr-
-awal es zenth kenet-
-tel meǵ kénéttet

Szkennelés, OCR



tegedeth az kerestfan
sebhetted meregĝel :
ees ecettel ytatal : ker-
-lek tegedet hog az
te zenth sebeyd legé-
-nek : en lelkemnek
orwossagy : amen
pf nf : Aue maria
O wram iesus
eristus ymadlak
tegedeth : coporsoba
hellieztettel myrr-
-awal es zenth kenet-
-tel meg kenettet

A korpusz mérete

Token–Type megkülönböztetés – példa(beszéd)

Mt 13,3-9

„Íme, kiment a magvető vetni. Amint vetett, némely szem az útszélre esett. Jöttek az égi madarak és fölcsipegették. Más mag köves talajba hullott, ahol nem volt neki elég föld. Gyorsan kikelt, mert nem volt mélyen a földben. Amikor azonban forrón tűzött a nap, elszáradt, mert nem volt gyökere. Ismét más szúrós bogáncsok közé esett. Amikor a bogáncsok felnőttek, elfojtották. A többi jó földbe hullott s termést hozott, az egyik százszorosat, a másik hatvanszorosat, a harmadik meg harmincszorosat. Akinek füle van, hallja meg.”

Token–Type megkülönböztetés 1., 2., 3.

11	,
10	.
6	a
3	volt
3	nem
3	az
2	mert
2	meg
2	hullott
2	esett
2	bogáncsok
2	Amikor
1	útszélre
1	és
1	égi
1	Íme

11	,
10	.
7	a
3	volt
3	nem
3	az
2	más
2	mert
2	meg
2	hullott
2	esett
2	bogáncsok
2	amikor
1	útszélre
1	íme
1	és

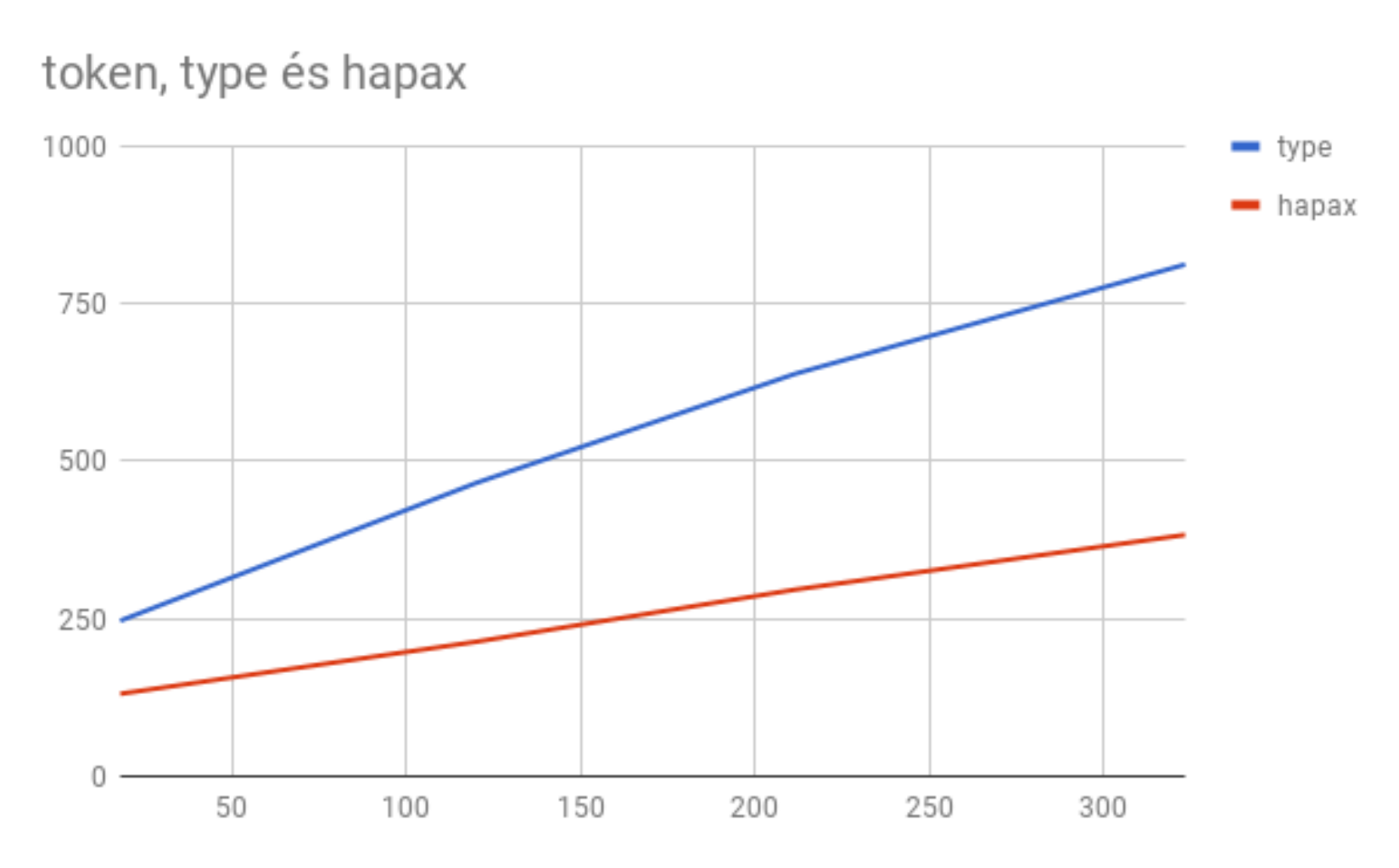
11	,
10	.
7	a
4	van
3	nem
3	föld
3	az
2	más
2	mert
2	meg
2	hull
2	esik
2	bogáncs
2	amikor
1	ő
1	útszél

A COBUILD mérete 1.

a Collins Birmingham University International Language Databank által készített szótárakhoz használt korpusz növekedése

	1987	1993	1995	1996
tokenek	18.000.000	120.000.000	211.505.963	323.302.789
típusok	247.069	475.633	638.901	812.452
hapax	131.299	213.684	296.436	383.356
nem-hapax	115.770	269.949	342.464	429.096
>10	43.579	104.201	134.942	164.963
>15	n.a.	n.a.	111.007	164.633

A COBUILD mérete 2.



Az MNSz2 mérete

	magyarországi	*szlovákiai*	*kárpátaljai*	*erdélyi*	*vajdasági*	összesen
sajtó	350,5	11,6	0,7	0,6	1,5	364,8
szépirodalom	77,0	2,3	0,4	0,8	0,2	80,6
tudományos	112,0	3,3	0,7	1,6	0,3	117,9
hivatalos	98,0	0,2	0,3	0,6	0,1>	99,0
személyes	300,3	-	0,4	0,4	0,1>	301,1
beszéltnyelvi	76,2	-	-	-	-	76,2
összesen	1013,9	17,3	2,5	3,9	2,0	1039,7

Magyar NE-korpuszok méretei

	LOC	*MISC*	*ORG*	*PER*	NEs	tokens	density(%)
Szeged NER	1,501	2,041	20,433	1,921	25,896	225,963	11.46
Crimi T-f-M	5,049	1,917	8,782	8,101	23,849	562,822	4.24
Crimi T-f-T	5,391	854	9,480	8,121	23,846	562,822	4.24

Korpuszannotáció

A nyelvi annotáció

a sztenderd szövegfeldolgozó lépések a modern korpuszoknál nagyjából ugyanazok:

- szegmentálás (tokenizálás, mondatra bontás)
- morfológiai elemzés
- morfoszintaktikai egyértelműsítés

Mi kell az annotációhoz?

- annotációs séma
  - elméleti nyelvészeti alapok lefektetése (pl. mi a tulajdonnév?)
  - címkekészlet
  - az annotáció formátuma (inline vagy standoff)
- annotációs eszköz
- az annotátorok száma $\rightarrow$ annotátorok közötti egyetértés mérése
- annotációs útmutató
- az annotáció minőségének ellenőrzése

Annotációs útmutató

- az útmutatónak egyszerre kell kellően kidolgozottnak és egyszerűnek lennie, hogy az annotátorok számára követhető legyen → ha nem így van, akkor az annotátorok magas hibaszázalékkal fognak dolgozni
- tartalmaznia kell az annotációs feladat leírását, az annotálandó nyelvi elemek felsorolását és példákat arra, hogy mit kell és mit nem kell annotálni
- minél magasabb nyelvi szintre megyünk, minél több szemantika van benne, annál képlékenyebb a feladat → bizonyos nyelvi jelenségek nehezen megfoghatók/formalizálhatók
- ha az útmutató nem elég egzakt, akkor az annotátorok elkezdik követni az intuíciójukat → a nem teljesen egyértelmű esetekben ez problémákat okozhat

NER annotációs útmutatók

- MUC-7 Named Entity Task Definition (Chinchor, 1997)
- ACE (Automatic Content Extraction) English Annotation Guidelines for Entities (Linguistic Data Consortium, 2008)
- Hunner project proposal és útmutató

Az annotáció formátuma

inline

```
<s><w>Ez</w> <w>egy</w> <w>mondat</w> <c> .</c>
<s><w>Meg</w> <w>a</w> <w>második</w> <c> .</c>
```

standoff

Ez
egy
mondat
.

Meg
a
második
.

XML

EXtensible Markup Language

egyfajta jelölőnyelv (markup language) → vannak más hasonlók: YAML, JSON, MD

Előnyei:

- mind ember, mind gép számára olvasható formátum
- támogatja a Unicode-ot
- szabványos és platformfüggetlen
- képes a legtöbb általános számítástudományi adatstruktúra ábrázolására

Hátrányai:

- szintaxisa elég bőbeszédű és részben redundáns
- nagyobb tárolási költség
- nincs lehetőség a dokumentum egyes részeinek közvetlen elérésére
- átfedő adatstruktúrák modellezése nehéz/lehetetlen

Standoff annotáció

- az eredeti dokumentumok sima szöveg fájlok maradnak
- az annotációk nem szövegközi tagek, hanem külön jelöljük úgy, hogy megadjuk, hogy az eredeti szöveg melyik karaktertartományára vonatkozik a címkézés, és hogy milyen címkét kap a szövegrészlet
- az annotálást teljesen különválasztjuk a használt feldolgozó eszközöktől, és közben minden információt megtartunk
- az átfedő és beágyazott annotáció is könnyen kezelhető

Beágyazott és átfedő annotáció

Beágyazott annotáció

<LOC><PERSON>Kossuth Lajos</PERSON>utca</LOC>

Átfedő annotáció

a Kossuth Lajos és a Petőfi Sándor utca sarkán

BIE1

[...] közölte Wolf László, az OTP Bank vezérigazgató-helyettese az MTI érdeklődésére.

közölte	O
Wolf	B-PER
László	E-PER
,	O
az	O
OTP	B-ORG
Bank	E-ORG
vezérigazgató-helyettese	O
az	O
MTI	1-ORG
érdeklődésére	O
.	O

BIE1

A szállásunk egy Balaton melletti kis üdülőfaluban, Zamárdiban volt.

A	B-NP
szállásunk	E-NP
egy	B-NP
Balaton	I-NP
melletti	I-NP
kis	I-NP
üdülőfaluban	I-NP
,	I-NP
Zamárdiban	E-NP
volt	O
.	O

Irodalom i

- Hunston, S. (2008). Collection strategies and design decisions. In: Lüdeling, A. and Kytö, M., editors, Corpus Linguistics. An International Handbook, pages 154–167. Walter de Gruyter, Berlin.
- Kugler, N. and Tolcsvai Nagy, G., editors (2000). Nyelvi fogalmak kisszótára. Korona, Budapest.
- McEnery, A. and Xiao, R. (2007). Parallel and comparable corpora: What are they up to? Translating Europe. Multilingual Matters.
- McEnery, T. (2004). Corpus Linguistics. In: Mitkov, R., editor, The OxfordH andbook of Computational Linguistics, pages 448–463. Oxford University Press, New York.
- Sinclair, J. (2005). Corpus and Text – Basic Principles. In Wynne, M., editor, Developing Linguistic Corpora: a Guide to Good Practice, pages 1–16. Oxbow Books, Oxford.

Irodalom ii

Javasolt olvasmányok:

- O'Keeffe, A. and McCarthy, M., editors (2010). The Routledge Handbook of Corpus Linguistics. Routledge, London and New York.
- Lüdeling, A. and Kytö, M., editors (2008). Corpus Linguistics. An International Handbook. Walter de Gruyter, Berlin.
- Szirmai, M. (2005). Bevezetés a korpusznyelvészetbe. A korpusznyelvészet alkalmazása az anyanyelv és az idegen nyelv tanulásában és tanításában. Tinta Könyvkiadó, Budapest.