1. MI A KORPUSZNYELVÉSZET?

1.1. Bevezetés

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

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

1.2. Mi a korpusz?

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

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

---

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1. táblázat: Relációs adatbázis

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

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

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

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

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

Overview for

"mind"

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

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

Search for Synonyms, ordered by estimated frequency of senses

Show glosses

Show contextual help

Search

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

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

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

1. ábra: WordNet keresés eredménye

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

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

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

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

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

1.3. A korpusz tervezése

1.3.1. A reprezentativitás

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

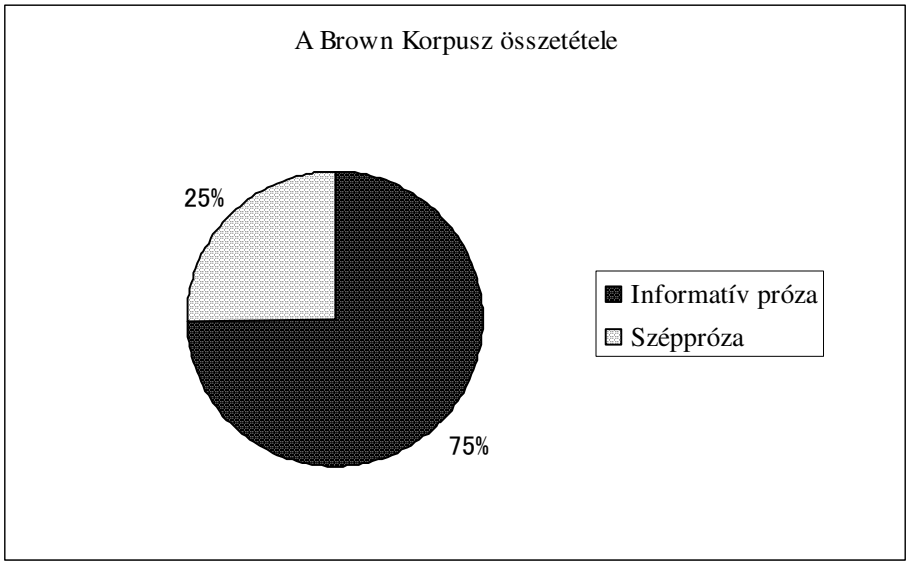
A legtöbb nyelvész eddig inkább az intuíciói és meggyőződése alapján választotta ki, nem pedig pontosan meghatározott elvek vagy statisztikai eredmények alapján, hogy mi kerüljön bele, és mi ne kerüljön a korpuszba. Jóllehet a szakirodalomban fontos szerepet tulajdonítanak annak, hogy a korpusz reprezentatív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

1.3.1.1. Mintavétel

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

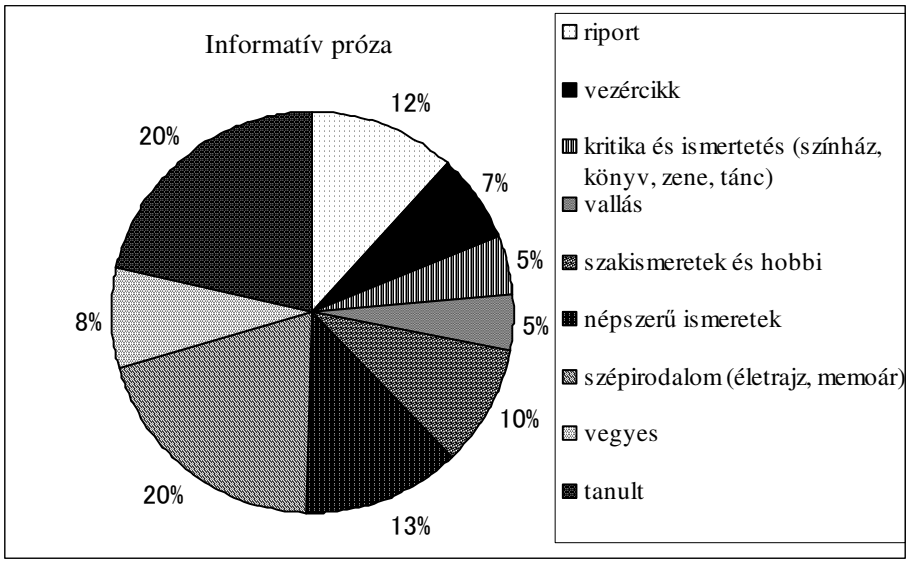
vagyunk tehát eldönteni, hogy a rendelkezésünkre álló regények közül egy adott típusú a korpusz hány százalékát tegye ki, pl. mennyi történelmi regény, sci-fi, életrajz vagy önéletrajz kerüljön a vizsgálandó gyűjteménybe. Szerepeljen-e vagy szerepelhet-e egy író összes műve, csak azért, mert könnyen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

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

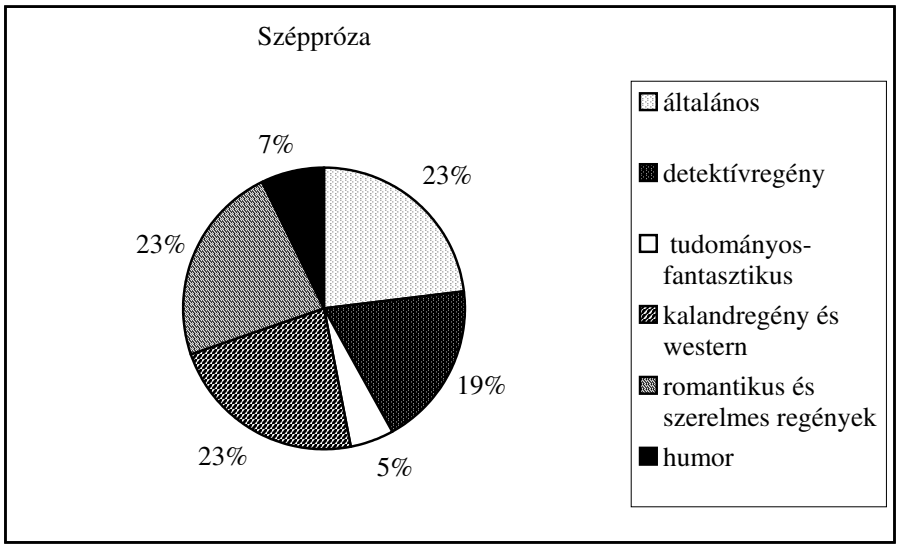


2. ábra: A Brown Korpusz összetétele

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



3. ábra: Az informatív próza alkategóriái a Brown Korpuszban



4. ábra: A szépprózа alkategóriái a Brown Korpuszban

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

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

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

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

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

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

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

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

1.3.1.2. A korpusz mérete

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

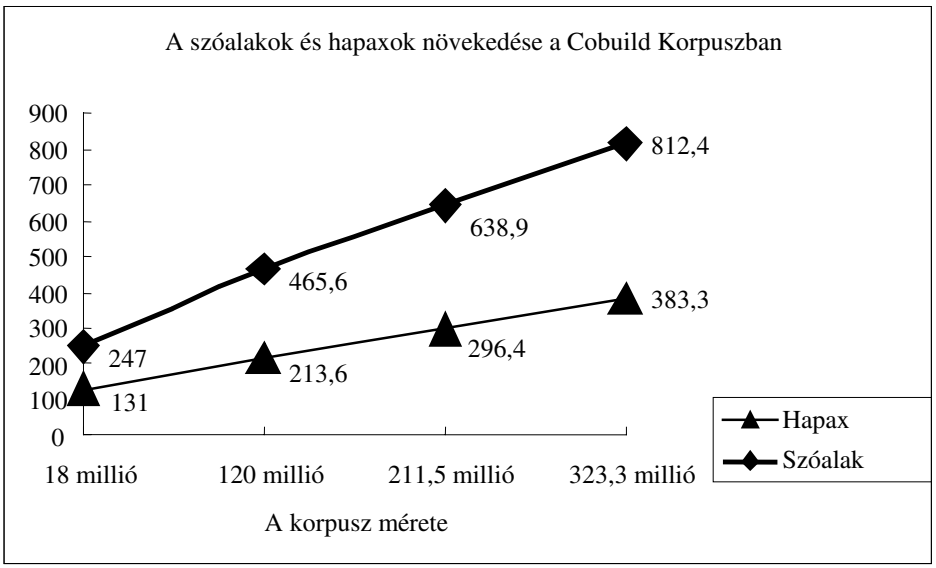
korpusz, a Brown Korpusz, az *Egri csillagok* mindössze hét és félszeresét, azaz 1 millió szót tartalmazott. A korpusz 500 darab, egyenként kb. kétezer szavas fájlból állt. Sok hasonló korpusz készült később ennek mintájára. Az angolban a méret megadásakor a szövegszó (running words) kifejezés használatos,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

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

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

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

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



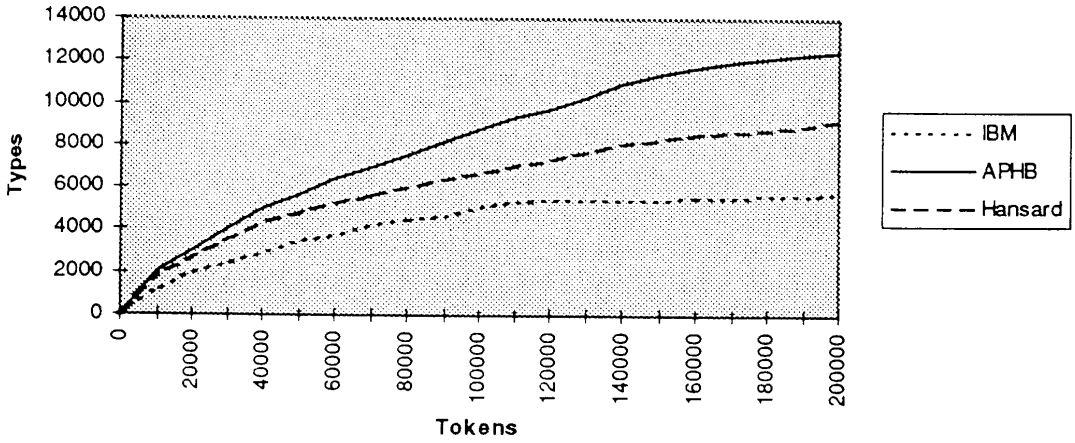
5. ábra: A típusok és a hapaxok növekedése a COBUILD Korpuszban

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

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

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

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



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

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

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

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

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

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

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

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

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

---

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

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

1.3.2. A korpuszok fajtái

1.3.2.1. A mintavétel módja szerint

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

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

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

1.3.2.2. A korpusz felhasználásának módja szerint

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

Általános korpusz

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

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

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

Speciális korpuszok

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

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

Összehasonlítható korpuszok (*comparable corpora*)

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

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

Párhuzamos korpusz (*parallel corpus*)

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

Nyelvtanulói korpusz

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

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

Pedagógiai korpusz

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

Történeti vagy diakrón korpusz

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

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

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

1.3.3. Jogi problémák

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

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

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

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

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

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

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

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

7. ábra: A Magyar Elektronikus Könyvtár figyelmeztető szövege

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

1.3.4. Átírás és annotáció

1.3.4.1. A beszéd átírása

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

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

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

8. ábra: Prozódikus átirat

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

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

1.3.4.2. A standard annotáció

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

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

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

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

4. táblázat: A Magyar Nemzeti Szövegtár alapkódjai

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

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

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

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

9. ábra: A BNC annotált szövege

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

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

10. ábra: A BNC szövege annotálás nélkül

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

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

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

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

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

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

5. táblázat: A Szerb Nyelv Korpuszának elemzett formája

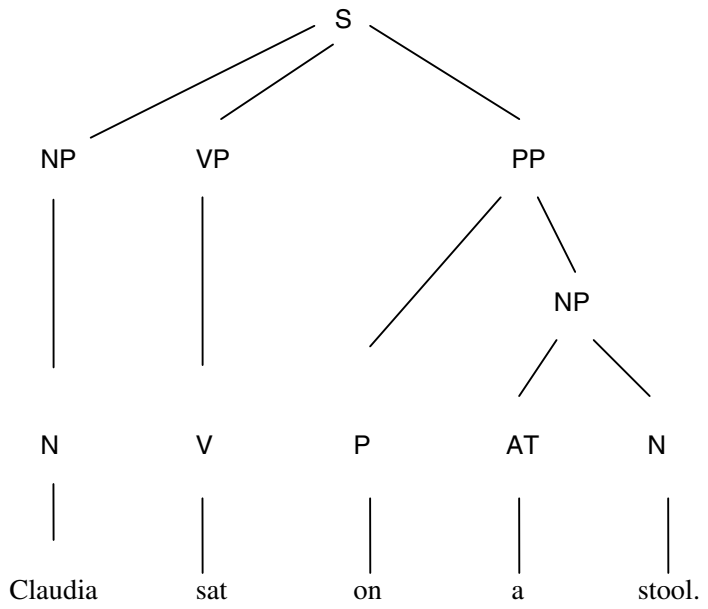
A mondattanilag elemzett korpuszok száma még kevesebb, mint a szófaji jelekkel ellátott korpuszoké. A mondattani elemzések két fajtája létezik: a „csontváz" elemzés (*skeleton parsing*) vagy „sekély" elemzés (*shallow parsing*) és a teljes elemzés (*full parsing*). Mint a megnevezésből sejthetjük, a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

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

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

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

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



12. ábra: Példa az ágrajzra

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

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

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

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

1.3.4.3. Speciális annotációk

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

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

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

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

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

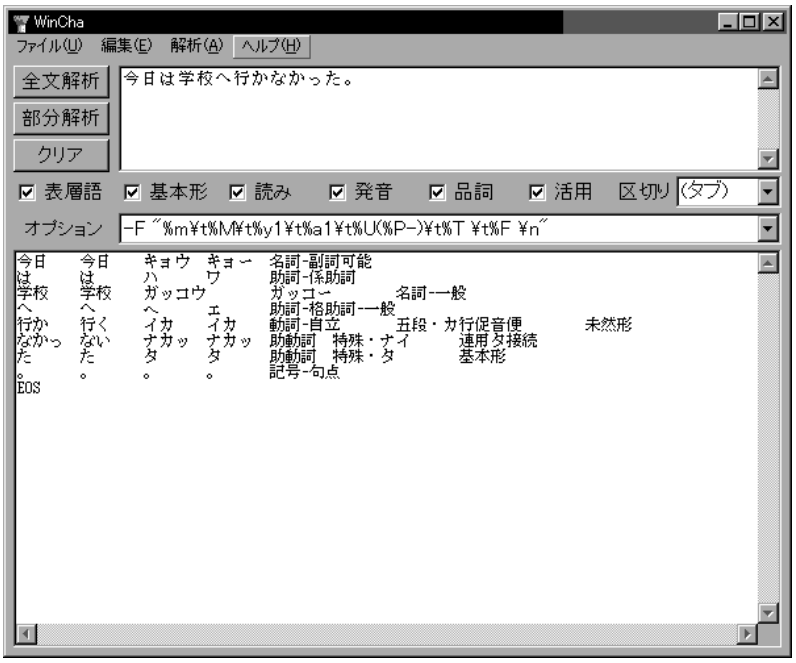
14. ábra: Diskurzus annotáció (Leech 1997: 13)

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

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

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

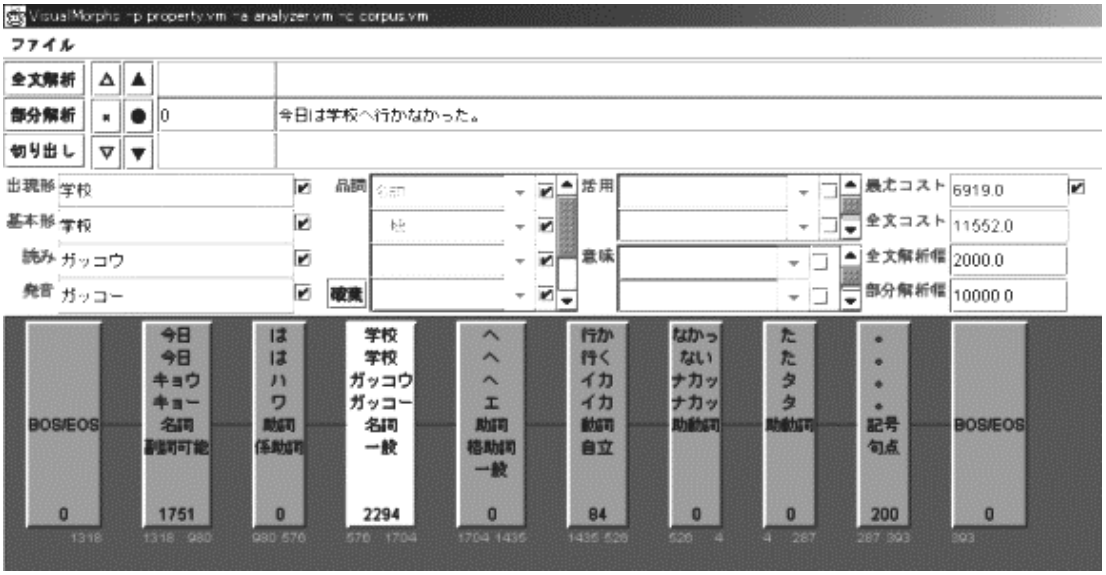
Mivel a korpusznyelvészet kialakulása során az angol nyelv elemzése volt szinte mindvégig a figyelem központjában, a „szabványos" annotációk is leginkább csak az angol nyelv elemzését tartják szem előtt (Oravecz & Dienes, 2002). Ezzel szemben minden egyes nyelvnek vannak olyan problémái, amelyeket lehetetlen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



15. ábra: Egy japán mondat morfológiai elemzése

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

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



16. ábra: Morfológiai elemzés vizuális elrendezésben

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

1.4. Összefoglalás

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

1. elektromos formában tárolt szövegek gyűjteménye;

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

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

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

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