TÓTH ÁGOSTON

Debreceni Egyetem, Angol–Amerikai Intézet
toth.agoston@arts.unideb.hu

Tóth Ágoston: Számolhatunk velük: szavaink egyenkénti és együttes előfordulási gyakoriságának mérése a lexikográfia szolgálatában
Alkalmazott Nyelvtudomány, Különkiadás 2017
doi:http://dx.doi.org/10.18460/ANY.K.2017.003

Számolhatunk velük: szavaink egyenkénti és együttes előfordulási gyakoriságának mérése a lexikográfia szolgálatában

Advances in information technology and quantitative corpus linguistics have had a great effect on lexicography. This paper focuses on word frequency and word co-occurrence frequency information. Word frequency is shown to have an effect on the headword inclusion lexicographical subtask and on compiling controlled defining vocabularies. Word frequency can also be meaningfully presented in dictionaries as a property of the headword. Frequency data may influence the ordering of senses within entries and the ordering of translation equivalents in bilingual dictionaries. Measuring first-order word co-occurrence frequency in corpora makes it possible to find and document collocations. Calculating second-order word co-occurrence information is a way of measuring the similarity of words, which also has lexicographical potential. Finally, a survey gives us insight into dictionary users' awareness of frequency-related dictionary features discussed in the paper.

1. Bevezetés

A szótárak szerkesztéséhez fokozatosan és hosszú idő alatt kifejlesztett eljárások az elektronikus számítógépek megjelenése után, a 20. század második felében sokat változtak. A szótári adatok számítógépes adatbázisokba kerültek, géppel feldolgozhatóvá váltak. Lehetségessé vált a szótárban tárolt információk típusainak bővítése, például a kiejtés hangzó anyagként tárolása, egyéb multimédiás tartalmak és hipertext használata. Az elektronikus szótárak megjelenésével megváltozott az információk elérhetőségének módja a szótárhasználók szempontjából is: a címszavak ábécé szerinti rendezése és abban való keresés, lapozgatás helyett az elektronikus szótárak keresőrendszerei az adatok szélesebb körében, a címszavak mögé látva, és sokkal gyorsabban elérhetővé teszik a szótári adatokat, miközben további eszközöket is kapunk (például inkrementális keresést a szólistában, helyettesítő karakterek támogatását, betűzési hibák automatikus javítását).

A cikk 2. fejezete egy tömör történeti áttekintésben ismerteti az informatika és a lexikográfia párhuzamos fejlődésének fontos találkozási pontjait. A 3. fejezetben azt vizsgáljuk, hogy a címszójelöltek gyakoriságának mérése hogyan segít a címszójegyzék összeállításában, valamint a címszó gyakorisága hogyan kezelhető és mutatható be szótári adatként. A 4. fejezet azt ismerteti, hogy a szócikkek elkészítése során milyen kvantitatív, szófrekvencia-alapú vizsgálati eszközökkel élhetünk, például a tanulói szótárak esetében a címszavak

értelmezését hogyan teheti érthetőbbé a gyakorisági adatok segítségével összeállított kontrollált definíciós szójegyzék („controlled defining vocabulary”) használata, és megvizsgáljuk, hogy a jelentésárnyalatok sorrendezésében milyen szerepe lehet a gyakoriságnak. Az 5. fejezet a szavak együttes előfordulási gyakoriságáról szól, egyrészt a szavak elsőrendű együttes előfordulásának méréséről, mely a szókapcsolatok kezelésében segíti a lexikográfust, másrészt a másodrendű együttes előfordulás jelenségéről, melynek elemzése a hasonló jelentésű szavak kezelésében hasznos. Az utóbbi téma olyan lehetőségeket is felvet, amelyekre vonatkozó lexikográfiai tapasztalat jelenleg még nincs, de a jövő szótárszerkesztési projektjeinek megtervezése során már *számolhatunk velük*. A cikk végén, a 6. fejezetben egy kérdőíves kutatást ismertetek, amely a cikkben is tárgyalt néhány jelenség, szótárírási fogás szótárhasználói ismertségét és elfogadottságát vizsgálja egyetemi hallgatók, többségükben leendő nyelvtanárok körében.

2. Szemelvények az informatika és a lexikográfia közös történetéből

A lexikográfiában az új technológiák korai adaptálása megszokott jelenség (Pajzs, 1990). A lexikográfusok és a kapcsolódó területek művelői, kutatói maguk is folyamatosan vizsgálják a megnyíló új lehetőségeket (pl. Prószéky, 2013).

A szavak előfordulási gyakorisága mint az egyes szavakat jellemző, gyakran kutatási (pl. nyelvtani, stilisztikai, bibliatudományi) célra felhasznált adat a gyakorisági szójegyzékek és konkordanciák formájában vált elérhetővé és széles körben kutathatóvá. Egy-egy gépi támogatás nélkül készített **manuális konkordancia** elkészítése jelentős kihívásnak számított, általában kiemelt művek és szerzők munkásságának az elemzését elősegítendő készítették el. Young csaknem 1700 oldalas konkordanciakötete Byron költészetéről (Young, 1965) például 25 éves munka eredménye, melynek elkezdésekor a digitális, elektronikus számítógépek még nem léteztek. A munka befejezésére azonban már nyilvánvalóvá váltak a számítógépek alkalmazásának előnyei, ebben az esetben az, hogy ez a feladattípus teljes egészében automatizálhatóvá vált. Young kitartó munkával, gépek nélkül fejezte be művét, de könyve előszavában maga is rámutat, hogy a kézzel készített konkordanciák ideje lejárt (Young 1965; hivatkozza: Howard-Hill, 1979: 81).

A konkordanciák a szavak szövegbeli előfordulásait környezetükkel együtt listázzák, és potenciálisan a feldolgozott művek eredeti szövegmennyiségének sokszorosát tartalmazzák, melyek nyomtatott formában történő kiadásának kezdetben nem volt alternatívája. A **szógyakorisági listák** – noha terjedelemre rövidebbek – ugyanúgy igénylik az adott szó összes előfordulásának szövegbeli megkeresését, valamint a szólista megfelelő rendezését is (ahol mind az ábécé szerinti, mind a gyakorisági lista külön elkészítendő egy nyomtatott kiadás

esetén). Az, hogy a szövegben egy-egy szó összes előfordulásának megkeresése automatizálható, a feladat időigényét napokra, órákra, majd másodpercekre rövidítette. Az 1966-ben megjelent *Szótártani Tanulmányok* (Országh, 1966) már külön fejezetben taglalja a gépi adatgyűjtés és adatfeldolgozás felhasználási lehetőségeit a lexikográfiában (Kelemen, 1966: 29-54), és ismerteti, hogy a gyakorisági szójegyzékek előállítása az adott technikai eszköztárral hogyan kivitelezhető.

A **szótári adatok számítógépes adatbázisban tárolása** és később a nyomtatott változat elektronikus, részben automatizált nyomda alá rendezése az informatika fejlődésével vált lehetővé. Az *Oxford English Dictionary* első számítógépes adatbázis alapú kiadása 1989-ben jelent meg (OED2) 7 év szerkesztőmunka után. A 20 kötetes, 22 ezer oldalas mű előkészítése az akkor újnak számító SGML jelölőnyelvet használva történt; a meglévő (nem elektronikus) lexikográfiai adatbázisok begépelése 120 adatrögzítő munkája volt, közben a lexikográfusok elvégezték a szótári adatbázis modernizálását és kiegészítését is. A kiadó 13,5 millió dolláros költségkeretről számol be ennél a projektnél (lásd http://public.oed.com/history-of-the-oed/). Az adatbázis elkészítése lehetővé tette, hogy 1992-ben a szótár elektronikus formában, CD-ROM-on is megjelenhessen, mellyel a 68 kilogrammnyi nyomtatott szótár nem egyszerűen egy új, kezelhetőbb és modernebb adathordozóra került fel, hanem egy olyan keresőeszközt is kapott, amellyel a címszavak és a korábban nem is kereshető szótári adatok széles köre vált nagyon gyorsan és egyszerűen kereshetővé.

A nyelvészeti célú alkalmazások előfeltétele az elemzendő **szövegek számítógéppel olvasható változatának előállítása**. A szövegeket gépeléssel (még korábban lyukkártyák lyukasztásával), később pedig optikai karakterfelismeréssel és annak utószerkesztésével tudták rögzíteni. Az 1960-as években kiadott, úttörő jellegű Brown korpusz hatalmasnak számított 1 millió szövegszavas méretével.

Az első **elektronikus korpusz alapján készített angol szótár** (COBUILD1) szövegkorpusza két módszerrel készült: gépeléssel, illetve optikai karakterfelismeréssel és annak manuális utószerkesztésével (pl. ’smali’ → ’small’, ’thc’ → ’the’, stb.). Renouf (1987) részletesen is ismerteti az eljárást, és ír a korpusztervezési elvekről is: 1960 utáni – azon belül is minél későbbi kiadású – szövegeket gyűjtöttek, nyelvváltozatokra vonatkozó kvótákat határoztak meg (70% brit, 20% amerikai, stb.), figyelték a szerzők korát (16 év fölötti, „felnőtt” nyelv), valamint a szövegek műfaját (a lírát és a drámát kizárták a „természetesség” kedvéért). A korpusz készítői beszélt nyelvi adatokat is felhasználtak, melyeket sztenderd angol ortográfiát használva jegyeztek le. Krishnamurthy (1987) arról számol be, hogy a szótár előállítása során 7,3 millió szavas korpusszal dolgoztak, melyből 1,3 millió szavas volt a beszélt nyelvi

rész. Más forrásokból, illetve szótári bevezetőkből tudjuk, hogy szótáraik későbbi kiadásai ennél nagyobb korpuszból készültek. A címszavakhoz konkordanciákat készítettek (ez az elektronikus korpuszból igen gyorsan megoldható volt), ezeket kinyomtatták, a lexikográfusok ezekkel dolgoztak. A szótárszerkesztők a konkordanciák alapján a címszavakhoz kézzel írt cédulákat készítettek, melyeket végül adatrögzítők mainframe nagygépeken rögzítettek.

Az adatok teljesen elektronikus tárolása és kezelése volt a fejlődés következő állomása, amit a személyi számítógépek elterjedése és tárolókapacitásuk fejlődése tett lehetővé. 1977-ben jelent meg az Apple II (az első sikeres otthoni számítógép) hobbi, szórakoztatási és oktatási célokra. Nagyobb adatbázisok kezelésére ez még nem volt alkalmas. 1981-ben megjelent az IBM PC (a mai asztali és hordozható gépeink őse) és a dBASE II adatbáziskezelő rendszer. Lexikográfiai célokra még egy ideig a vállalati mainframe számítástechnika megfelelőbb volt (ahogyan azt a COBUILD példáján is láttuk), de az 1990-es évekre már széles körben elterjedtté váltak a megfelelően kiépített személyi számítógépek is. Ezek fokozatos irodai elterjedése és otthoni megjelenése tette lehetővé a **teljesen elektronikus munkavégzést**.

Egyre több szöveg született meg elektronikus változatban, és az elterjedőben lévő Interneten ezeket nagy távolságokra is hatékonyan lehetett szállítani, terjeszteni. A számítógép-használat gyors terjedése, valamint a HTTP és HTML szabványokra épülő, hipertext alapú **World Wide Web** megjelenése a 20. század utolsó évtizedében minőségi változást is hozott. Rengeteg szöveg jelenik meg a weben napról-napra, így akár milliárdnyi szövegszót tartalmazó korpuszokat gyűjthetünk automatizált módszerekkel (lásd pl. GloWbE, NOW).

Szavaink egy kisebb része igen gyakori, emiatt jól megfigyelhető kis korpuszban is, nagyobb részük viszont ritka, és nagyobb korpuszt igényel akár egyetlen előfordulás megatalálása is. A ritka szavak egyik forrása a nyelvi produktivitás, a szóalkotás lehetősége. A kevésbé gyakori jelenségek tanulmányozásában a nagy és műfajilag változatos korpuszok segítenek. Fontos a korpuszok folyamatos karbantartása, bővítése is, hogy a nyelv változását követni, „monitorozni” tudjuk, és a megjelenő új jelenségekről információkkal rendelkezzünk.

A hardveres erőforrások folyamatos és gyors fejlődését (a feldolgozási sebesség változását, az operatív memória növekedését, a perifériák tárolókapacitásának és sebességének a növekedését, a hálózatok terjedését és sebességük felgyorsulását) kiegészíti a szoftverek fejlődése. Az adatbázisokat és korpuszokat kényelmesen, leggyakrabban a weben, saját adatbázis-kezelő rendszerükön keresztül érjük el. A korpuszokból konkordanciákat készíthetünk, alapvető statisztikai adatokat állíthatunk elő. A korpuszokkal való munkát nyelvtechnológiai fejlesztések is támogatják, például hasznos, ha a lekérdezések összeállításánál a szavakat lemma formájában is megkereshetjük, valamint ha szófaji és egyéb nyelvtani kategóriákkal is tudunk dolgozni. A magyar nyelvben

az alaktani elemzés erős eszközt ad a korpusz felhasználójának. Az elemzők közül a legismertebbek a *Humor* (Prószéky–Kis, 1999; Novák, 2003) és a *Hunmorph/morphdb.hu* rendszer (Trón et al., 2006). A nemzetközi piacra kitekintve a *Sketch Engine* szoftvereszköz (Kilgarriff et al., 2014; https://www.sketchengine.co.uk/) széles palettán kínál korpuszfeldolgozási eszközkészletet. A választott korpuszokból a szavakra jellemző statisztikai adatokat ki tudja nyerni (a szavak egyenkénti és együttes előfordulási gyakoriságát is beleértve), kollokációjelölteket ad, konkordanciákat készít, és automatizált módon (szabályalapú és statisztikai módszerekkel elkészített) szóskicceket („word sketch”) készít, melyekben az adott szavak tipikus használatát, grammatikai és/vagy más szavakkal jellemezhető tipikus környezetét automatikusan dokumentálja. A szoftvert kiadó, Adam Kilgarriff által alapított cég a vevői közt tartja számon a Collins, Cambridge, Le Robert, Macmillan és Oxford szótárkiadókat.

Azon túlmenően, hogy a nyelvtechnológiai megoldásokat korpuszfeldolgozási eszközökként felhasználhatjuk szótári adatok gyűjtése közben, ezek az eljárások **elektronikus szótárak keresőrendszerében** közvetlenül is megjelenhetnek. Ezen a ponton a helyesírás-ellenőrzést és -tanácsadást emelem ki példaként. A *Merriam–Webster* online szótár (MW online) az angol *damage* szó magyar nyelvtanulók által fonetikai okok miatt gyakran használt, hibás *demage* változatának begépelésekor 17 javítási javaslatot ad, köztük első helyen a *damage* szóval, ezáltal a szótár egyrészt felhívja a felhasználó figyelmét a helyesírási hibára (ilyen módon is oktatási segédeszközzé válva), másrészt a javaslatok listájáról elérhető a helyes címszó.

3. A címszavak gyakorisága

3.1 A szógyakoriság szerepe a címszójegyzék összeállításában

Kilgarriff az LDOCE3 szerkesztésében részt vevő lexikográfusoktól kérdőíves felmérés keretében adatokat gyűjtött arról, hogy melyik szótárszerkesztési részfeladat milyen nehézségű a megítélésük szerint (Kilgarriff, 1998). A felmérésben a címszójegyzék összeállítása nem tartozott a nehéz feladatok közé (a 13 részfeladatból a 9. helyen végzett), de ennek az is az oka, hogy ezzel kevesebben és rövidebb időtartamban foglalkoztak; ez inkább a stratégiai tervezés és előkészítés része.

A szótár terjedelme az egyik alapadat, mely a szótár kategorizálását is meghatározza. A címszavak mennyisége a szótár előállításához szükséges munkamennyiséget is befolyásolja.

A nyomtatott szótáraknál megszokhattuk, hogy különböző terjedelmű szótárakat használunk különböző célra. A legteljesebb címszó- és jelentésárnyalat-készletű nagyszótárak a szakszókincs egy részét is tartalmazzák az elterjedt szakterületeken, bizonyos mértékig a szakfordítói munkát is

támogatva. A középszótár méretbe ennél kevesebb adat való, főleg a szakszavak tekintetében, ellenben a köznyelv gyakori szavait, valamint a fontos jelentésárnyalatokat ennek a szótárméretnek is tartalmaznia kell. A kisszótárakba a mindennapi iskolai munkához, vagy például utazáshoz szükséges szókincs kerül. Bár elvileg a nagyobb címszómennyiség a szótár hasznosságát növeli, de a nyomtatott szótárak esetében egyben a hordozhatóság és használhatóság (keresési sebesség) rovására is megy.

Arról, hogy a szakszókincs az általános szótárakban hogyan és miért jelenik meg, beleértve a középszótár méretet is, Prószéky (2011) a szerkesztői, kiadói szempontokat is ismertetve ír. A gyakorlatban is rendszeresen találkozunk szakszavakkal a nagyszótárnál kisebb méretű, általános szótárakban is. Például a Magay–Országh magyar–angol kéziszótár (MA) is tartalmaz 29 „*gabona*-alapú” címszót (többségük szóösszetétel *gabona*- előtaggal); ezek a szótár egy oldalának körülbelül a negyedét foglalják el. E szavak között szerepelnek az MNSZ2 korpuszban jól reprezentált *gabonatermés* és *gabonapehely* szavak (595, illetve 442 előfordulás a lemmákra), de ugyancsak megtaláljuk köztük pl. a *gabonarozsda* és *gabonaüszög* szavakat is, amelyek a korpusz több, mint 1 milliárd szavából 23, illetve 4 alkalommal jelennek csak meg. Azt, hogy e szakszavak középszótári szerepeltetése miért volt a szerkesztők számára fontos, egyáltalán ez mennyiben volt tudatos döntés a részükről, nem tudom megmondani, de a köznyelvet jól – megtervezett módon – reprezentáló korpuszok segítségével a címszavakról eldönthető, hogy középszótár méretben a beválasztásuk indokolt-e.

A szótárak fejlődése, változása ugyanakkor a különböző terjedelmű szótárak közötti műfaji határokat is elmossa. Az elektronikus szótárak esetében a terjedelem nem függ össze szembetűnő módon sem a hordozhatósággal, sem a keresési sebességgel. Általában a minél nagyobb terjedelem az előnyösebb, marketing szempontból is.

Mindez összefügg azzal a kérdéssel is, hogy a szótárak árát elfogadható szintre sikerül-e csökkenteni, például éppen úgy, hogy a projekteket a kevésbé fontos költségelemektől megtisztítjuk. Bár a költségeket akár címszóról-címszóra is csökkenthetjük valamelyest, a kihívás óriási: egy nemzetközi, nagy mintát vizsgáló (n=684) kérdőíves kutatásban (Koplenig–Müller-Spitzer, 2014) a válaszadók csupán 16%-a nyilatkozott úgy, hogy a szótárakért fizetne. Az elsöprő többség az ingyenes szótárakat tartja elfogadhatónak, de a reklám-alapú finanszírozást nem veti el. Több webszótárban találkozunk is hirdetésekkel (még az OLD online-ban is), és létezik olyan angol egynyelvű szótár (a Macmillan Dictionary, lásd MAC online), ami hirdetésekre épít, és nyomtatott szótárként már nem jelenik meg. Persze az angol szótárak különleges helyzetben vannak, a kereslet hatalmas, a kínálati oldalon pedig komoly versenyhelyzet van. Visszatérve a gyakorisági adatokhoz, egyben a fentieket összefoglalva: a címszójegyzék gondos tervezése csökkenti a költségeket, a szükséges

címszavakra tudunk koncentrálni, elektronikus szótár esetén pedig a bővítés fokozatosan, akár közösségi módszerekkel is megtörténhet.

Néhány kiadó az online szótárait rendszeresen frissíti. Az *Oxford Learner's Dictionary online* szótár (OLD online) 2017 tavaszán, 2016 tavaszán és 2015 decemberében bővítette címszólistáját (egyszerre nagyobb mennyiségű címszóval); az új szavakról szóló hírt a szótár a címlapján hirdeti. A *Cambridge Dictionary online* szótár (CAD online) frissítése gyakoribb, bár egyszerre kevesebb címszót érint (https://dictionaryblog.cambridge.org/category/new-words). A *Longman Dictionary of Contemporary English online* (LDOCE online) szótár általában is nagyon kevés információt nyújt az elektronikus szótárváltozatról, új címszavakat sem említ. Szúrópróbaszerű ellenőrzéssel a *brexit* szó a cikk megírásakor (kb. 1,5 évvel a brexit népszavazás után) még nem szerepel benne, miközben a versenytársak többségénél már címszó. Ugyanakkor például a *selfie* neologizmust már tartalmazza ez a szótár is.

A gyakorisági listák rendszeres generálása és vizsgálata mellett a szótárak felhasználói által keresett – de az adatbázisból hiányzó – szavakat is fel lehet használni a címszójegyzék karbantartására. A meghiúsuló keresések megfigyelésén keresztül automatikusan javasol szótárbővítést a Lexitron modell (Trakultaweekoon et al., 2007).

A MAC online „Open Dictionary" programjában (https://www.macmillandictionary.com/open-dictionary/) a felhasználók közvetlenül ajánlhatnak címszavakat és meglévő címszavakhoz jelentésárnyalatokat. Ilyenkor egyúttal megadják az általuk javasolt értelmezést, és példamondatokat is küldhetnek. A projekt 2009 óta aktív; 8 év alatt 4000 új címszót vettek fel a szótárba a szerkesztők ennek segítségével. A javaslatot tevő felhasználó nevét a szótári adatbázis megőrzi. Előfeltétel, hogy Google kereséssel független forrásban is megtalálható legyen az adott szó.

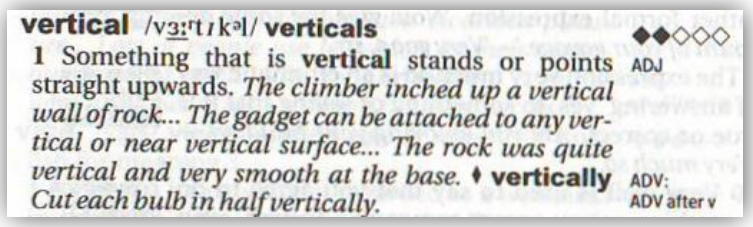
A Collins kiadó elektronikus szótárába (Collins online) szintén lehet címszójavaslatokat beküldeni a szótár blogján keresztül (https://www.collinsdictionary.com/word-lovers-blog, „Submit a word"), a jelentésárnyalatokra, értelmezésekre vagy egyéb szótári adatra vonatkozó megjegyzéseinket pedig elhelyezhetjük közvetlenül a szócikkek alatt. Ezek a funkciók regisztrációt igényelnek, és a kiadó részéről moderációt kapnak.

3.2 A címszógyakoriság mint szótári adat

A szavak előfordulási gyakorisága nem kizárólag a szótárszerkesztő számára fontos adat, hanem a szótárhasználó számára is értékes.

Az angol egynyelvű szótárak közül a korpusz alapú COBUILD szótárakban a frekvenciaadatok megadása volt az egyik új szolgáltatás. Az 1. ábra jobb felső sarkában látható grafikus skálán adja meg a szótár a címszó gyakorisági adatát (itt a COBUILD2 kiadásból vett példán). A szótár bevezetőjéből a

szótárhasználó megtudja, hogy az első gyakorisági sávba (amit grafikusan 5 kitöltött rombusszal jelölnek) a leggyakoribb kb. 700 szó került (pl. *the, and, of, like, go, paper*), a másodikban hozzávetőlegesen 1200 szó található (pl. *argue, bridge, danger, female*), és a két sáv együtt nagyjából „az angol nyelvhasználat" (feltehetően a korpusz tokenmennyiségének) 75%-át fedi le (COBUILD2, xiii), ezért „ezen szavak fontossága nyilvánvaló" (ibid., saját fordítás). A harmadik sávba 1500, a negyedikbe 3200, az utolsóba kb. 8100 szó került; az öt gyakorisági sáv együtt az írott és beszélt nyelvhasználat 95%-át fedi le a szerkesztők adatai szerint. A gyakorisági skálán nem ábrázolt szavakat a bevezető explicit módon kevésbé fontosként jellemzi, azonban ezekben az esetekben is megnyugtatja a szótárhasználót, hogy kizárólag azokat a szavakat vették fel a címszavak közé, amelyekkel korpuszadataik alapján a használó a legnagyobb valószínűséggel találkozik (ibid.).

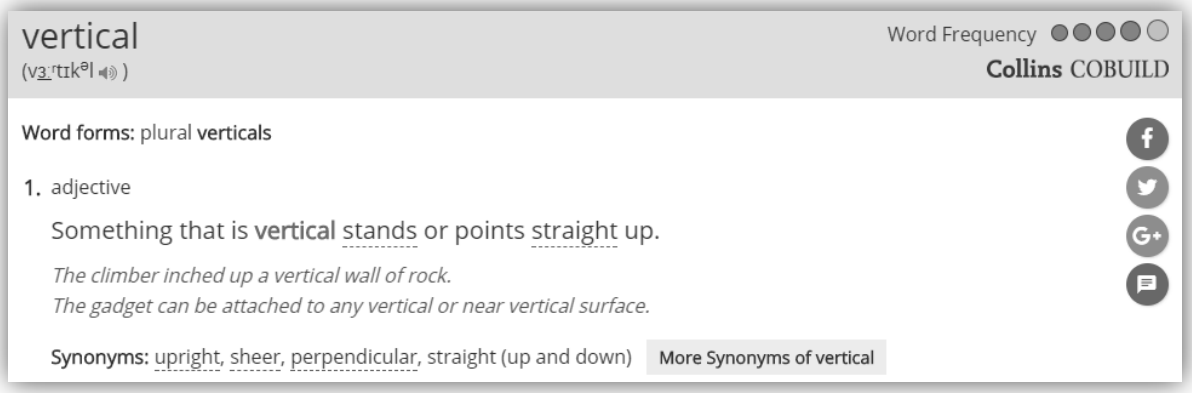

vertical /vɜːtɪkəl/ **verticals** ◆◆◇◇◇

1 Something that is **vertical** stands or points straight upwards. *The climber inched up a vertical wall of rock... The gadget can be attached to any vertical or near vertical surface... The rock was quite vertical and very smooth at the base.* ♦ **vertically** *Cut each bulb in half vertically.* ADJ ADV: ADV after v

1. ábra: A *vertical* szó melléknévi és határozószói használata a Collins Cobuild English szótárban (COBUILD2)

Saját mérésemben, egy erre a célra kiválasztott 100 millió szavas Wikipedia pillanatfelvételben (ami természetesen egy más jellegű korpusz, mint amivel a szótárszerkesztők dolgoztak) a *vertical* szó összesen 2032 alkalommal fordult elő, ezzel az 5116. helyen végzett a fellelt, hozzávetőlegesen 1 millió szótípusból.

A 2. ábrán láthatjuk, hogy a COBUILD sorozat leszármazottja, a Collins online szótár ugyanezt a szót hogyan mutatja be. A két bejegyzés hasonló, de az 5 fokozatú skála beosztása (és ezzel a *vertical* szóhoz megadott gyakorisági adat) jól láthatóan megváltozott.



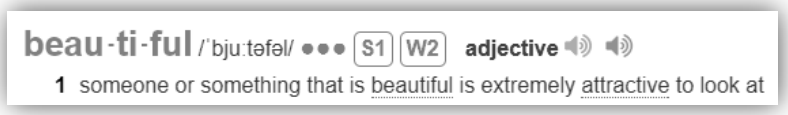
2. ábra: A *vertical* szó melléknévi használata a Collins online szótárban
(https://www.collinsdictionary.com/dictionary/english/vertical)

A *Macmillan* online szótár (MAC online) szintén közli a címszavai gyakorisági adatait. Az általuk használt korpusz 7500 leggyakoribb címszavát pirossal jelenítik meg, valamint egy 3 fokú skálán (grafikus módon, csillagokkal) a gyakoriságukat is megkapjuk. A 3 gyakorisági sáv mindegyike 2500-2500 szót tartalmaz. A 3. ábrán látjuk példaként a *vertical* címszóhoz tartozó szócikk egy részletét.

3. ábra: A *vertical* szó melléknévi használata MAC online szótárban,
a címsor végén a 2. frekvenciasávot jelző gyakorisági adattal
(https://www.macmillandictionary.com/dictionary/british/vertical_1)

A Longman kiadó az LDOCE3 óta több célra is használja a szógyakorisági adatokat (Summers, 1996); ez egyik ilyen cél a címszavak gyakoriságának ábrázolása szótári adatként. A 2003-as nyomtatott kiadásban a leggyakoribb 3000 címszót pirossal szedték, valamint az S1, S2, S3, W1, W2, W3 címkékkel közvetlenül a címszó után jelezték, hogy melyik regiszterben (beszélt vagy írott angol) melyik gyakorisági tartományban található az adott szó. Az S1 címkével a beszélt nyelvi korpusz leggyakoribb 1000 szavát, a W3 címkével az írott nyelvi korpusz 2001-3000 gyakorisági tartományát jelölték. Természetesen

előfordulhat, hogy ugyanaz a címszó az egyik regiszterben más gyakorisági tartományba kerül, mint a másikban, például a *beautiful* szó S1, W2 címkéket kapott. Az is megtörténhet, hogy az adott szó valamelyik nyelvváltozatban gyakori, a másikban nem (pl. *crash* [n] S3). A nyomtatott szótárhoz írt bevezető ezt a szolgáltatást pontosan leírja, valamint a tanárok és tanulók szempontjából „nagyon népszerűnek" titulálja. Az ingyenes LDOCE online webszótárban szintén látjuk a gyakorisági címkéket (4. ábra).



4. ábra: A *beautiful* címszó az LDOCE online kiadásában
(https://www.ldoceonline.com/dictionary/beautiful)

A beszélt és írott nyelv gyakorisági adatainak megkülönböztetése olyan szolgáltatás, amit a legnagyobb vetélytársak közül egyik sem ajánl fel, ugyanakkor az LDOCE 3000 leggyakoribb szavához képest a MAC online szótárban 7500 szóhoz, a COBUILD2 szótárban pedig a fent ismertetett módon kb. 14700 szóhoz kapunk gyakorisági adatot, ami a haladó nyelvtanulók igényeinek jobban megfelel, hiszen az általuk keresett szavak a 3000-es tartományon kívül esnek.

4. A gyakorisági adatok felhasználása a szócikkek elkészítése során

4.1. A kontrollált definíciós szójegyzék („controlled defining vocabulary", CDV)

Az értelmezések érthető megfogalmazása, az értelmezésekben felhasznált szókincs megfelelő megválasztása a tanulói szótárak esetében különösen fontos. A *Longman Dictionary of Contemporary English* szótárakban (pl. LDOCE3, LDOCE online) az értelmezések szavait egy körülbelül 2000 szótípusra korlátozott listáról („Longman Defining Vocabulary", LDV) választják ki a szótárszerkesztők. Ennek a definíciós szójegyzéknek az összetételét, méretét, változását Xu (2012) elemzi; saját mérési adatai szerint az angol nyelv 1000 leggyakoribb szócsaládja alkotja a szójegyzék kb. 45 százalékát (Xu, 2012: 370). A Longman gyakorlatának előzménye a *New Method English Dictionary* (NMED; kb. 24 ezer címszó, 1490 szótípussal értelmezve). Követői között pedig ott vannak a következő ismert szótárak is (Xu, 2012):

- a nagy vetélytárs *Oxford Advanced Learner's Dictionary*, mely 1995-től (OALD5) kezdődően használja saját 3000-3500 szavas CDV-jét,
- a *Cambridge Advanced Learner's Dictionary* kiadásai, illetve előzményeként a CIDE, 1995-től, kevesebb, mint 2000 szavas CVD-vel, és
- a *Macmillan English Dictionary for Advanced Learners* szótárak (MED1-től, 2002, kevesebb, mint 2500 szavas CDV-vel).

A CDV-k használatának előnye, hogy a nyelvtanuló számára könnyen érthető értelmezéseket eredményez, így azok további szótározására nincs szükség. Példaként álljon itt az angol *microphone* szó értelmezése a CDV alapú LDOCE online (tanulói) és a hagyományos módon megfogalmazott MW online (általános értelmező) szótárakból:

- LDOCE online: "a piece of equipment that you speak into to record your voice or make it louder when you are speaking or performing in public" (https://www.ldoceonline.com/dictionary/microphone)
- MW online: "an instrument whereby sound waves are caused to generate or modulate an electric current usually for the purpose of transmitting or recording sound (such as speech or music)" (https://www.merriam-webster.com/dictionary/microphone)

Láthatjuk, hogy az LDOCE változat egyszerű, lényegre törő, bár nem túl részletes (viszont kapunk az értelmezés mellé egy fényképet is); az MW értelmezése pedig már a fizikai hátteret is feszegeti ('a hanghullámok az áramerősséget modulálják…'), már-már szakkönyvi jelleggel.

A könnyebb érthetőség nem lebecsülendő előny, hiszen a szótárhasználó számára kevés kellemetlenebb helyzet van annál, mint amikor az értelmezést sem tudja értelmezni. A CDV-k használata pedig általában is segít elkerülni azt a helyzetet, amikor egy ritka szót egy másik nehezen érthető szóval (pl. egy szakszóval, vagy a címszó egy még ritkább szinonimájával) próbálnánk megmagyarázni. Elvi hátránya az értelmezések potenciálisan túlzott leegyszerűsítése, a pontosság csökkenése.

4.2 A jelentések (jelentésárnyalatok) sorrendezése

A szótárakban a jelentés egy szerkezeti egység a szócikken belül, és a lexikográfusok számára nem prioritás az, hogy ennek kialakításában bármilyen konkrét nyelvészeti modellt kövessenek. Cruse (2011) például a poliszémia kérdéskörében számos jelenséget dokumentál (az elkülöníthető jelentések észleléséről, a jelentés megváltozásáról kontextus függvényében, metaforákról, stb.); a szótárírás gyakorlatában ezek a kérdések részben máshogyan merülnek fel, sőt nagyobbrészt fel sem merülnek. A jelenségek osztályozása, címkézése és magyarázata kevésbé fontos, helyette elsősorban a korpuszban talált jelenségeknek a szótárhasználó számára érthető és hasznos megragadásával kell értelmezést vagy fordítási ekvivalenst adni, a szótárhasználó pedig saját nyelvi intuícióit és a világról gyűjtött tudását fel fogja használni a szótárból kapott adatok értelmezésében. A poliszémia és a homonímia közötti különbségtétel persze a szótár szerkezetére is kihat, de általánosságban elmondható, hogy a nyelvészeti kategóriák finom szövete, a jelenségek ezeken keresztüli láttatása és magyarázata itt másodlagos.

Lexikai szemantikai modellekből nem levezethető, mégis a gyakorlatban jól működő módon a többszörösen poliszém szavak esetében jó megoldás lehet a több jelentésárnyalatot lefedő, generikusabb jelentéssel kezdeni a címszó értelmezését, így a szótárhasználó sokszor már a szócikk olvasásának elején értékelhető információhoz jut, illetve látja, hogy jó helyen keres-e, vagy a szócikk teljesen másik részén kell folytatnia a keresést (esetleg másik szócikkben, homonímia esetén).

Sok esetben követett módszer az, hogy a jelentéseket a szócikken belül vezérszavak („guidewords", „mini-definitions") köré szervezik, melyek segítik a felhasználót a navigációban, keresésben. A vezérszavak megfelelő megválasztása egy pragmatikus és címszóra szabott eljárás; a címkék rendszere néha teljesen heterogénnek tűnik, mégis nagyon hasznos fogódzókat ad. Az *Oxford Learner's Dictionary* (online változatban is) például a következő vezérszavakat használja az angol *throw* ige különböző jelentéseinek elrendezésére: „with hand", „put carelessly", „move with force", „part of body", „make somebody fall", „into particular case", stb.

A jelentések a szócikken belül a szerkesztők által megválasztott sorrendben, akár hierarchikus szerkezetben jelennek meg. A COBUILD1 szótár szerkesztésekor követett gyakorlatról Moon (1987) ír. Rámutat, hogy a jelentések sorrendjének meghatározása közben több – egymásnak részben ellentmondó – szempontot kell vizsgálni, ezek egyike a jelentés gyakorisága. Ismert például, hogy a *summit* címszó esetében a hagyományosan előre sorolt „hegycsúcs" jelentés helyett a „csúcstalálkozó" használat került előbbre a szerkesztők korpuszalapú kutatásának gyakorisági adatai alapján. A szavak jelentéskategóriáit „főleg gyakorisági" sorrendben kéri bemutatni a kategóriákra vonatkozó COBUILD szerkesztői útmutató is (Krishnamurthy, 1987: 70).

Egyúttal azonban idézzük fel azt is, hogy a COBUILD projektben a lexikográfusok nyomtatott konkordanciákkal dolgoztak (lásd 2. fejezet), így a konkordanciaadatok kategorizálása és a jelentések megszámlálása manuális munkával történt. Ezen az adatok nem jelentek meg elektronikusan: abban a fázisban, amikor a lexikográfus által kézírással rögzített rekordokat számítógépen eltárolták, a kategóriák, jelentések hierarchiája és sorrendje már kialakult, a jelentésekre vonatkozó frekvenciaadatok pedig nem szerepeltek a rögzítendő adatok közt (v.ö. Clear, 1987: 48).

Az ismertebb angol egynyelvű szótárak közül az LDOCE3, illetve az LDOCE online a jelentések sorrendezésének elsődleges rendező elveként emeli ki a jelentések gyakoriságát. A szótári bevezető a *lookout* szó szócikkét hozza példaként, amiben egy frazális egység került az első helyre („*be on the lookout for somebody/something"*), a címszó önálló használatának magyarázata pedig későbbre került az alacsonyabb relatív gyakoriság miatt.[1]

Lew (2013) a jelentések sorrendbe állításához a következő négy elvet sorolja fel:

- a jelentések gyakoriságának elve;
- kronológiai sorrend: a *Merriam–Webster's Collegiate Dictionary* szerkesztői például a kronológiai sorrendről – az értékesítői csatornákon befutó – frekvencia alapú rendezést kérő felhasználói visszajelzések ellenére sem mondtak le, lásd (Morton, 1994: 82);
- jelöltség („markedness"): szociolingvisztikai és pragmatikai szempontból semleges, nem jelölt jelentések előbbre sorolása;
- logikusság: szótárspecifikus intuitív eljárások, pl. a jelentésárnyalatok központi jelentések („core senses") köré rendezése és a periférikus jelentésárnyalatok ezekhez viszonyított meghatározása azok speciális vagy generikus jelentésárnyalataként.

Ahogyan azt Morton (1994: 82) megjegyzi, a jelentések gyakoriság szerinti rendezése azért hasznos, mert a szótárhasználó is nagyobb valószínűséggel találkozik a gyakoribb jelentésekkel, így a szócikken belüli keresés felgyorsulhat. Ugyanakkor rámutat, hogy előfordulhat az is, hogy a szótárhasználó éppen a ritkább, kevésbé gyakori jelentéseket keresi, azok ismeretlensége okán.

A jelentések csoportosítása és bemutatása tehát komplex feladat, amiben a frekvencia elve nem mindig érvényesül. Ha elfogadjuk azt, hogy a szótárhasználó számára a jelentések korpuszban megfigyelt gyakorisága hasznos információ (akár a szó egyik tulajdonságaként, akár csak a szócikken belüli navigáció céljából), akkor a jelentés gyakoriságának rendezési elvként való használata helyett (vagy mellet) megfontolandó egy jelentésgyakorisági skála bevezetése a címszófrekvencia ábrázolásához hasonlóan.

Ugyanakkor a jelentések gyakoriságának meghatározása költséges munkafolyamat, mert ez az adat nem nyerhető ki automatizált módszerekkel a korpuszból, hanem manuális lexikográfiai munka eredménye. Miután a jelentések (jelentésárnyalatok) listája elkészül, a konkordancia sorait kategorizálni kell, és ezután állítható elő a jelentések gyakoriságának sorrendje. A jelentések megbízható automatizált klaszterezése és osztályozása sajnos a jelentésegyértelműsítés („word sense disambiguation", WSD) területén végzett több évtizednyi intenzív kutatómunka után sem olyan minőségű, ami a lexikográfus számára hasznos.

4.3 A fordítási ekvivalensek gyakorisági adatai

A kétnyelvű szótárak a szó jelentését fordítási ekvivalensek bemutatásával adják meg. Amennyiben a címszó nyelve a szótárhasználó anyanyelve és az ekvivalensek idegen nyelvűek, úgy különösen fontos, hogy a szótár tartalmazza azokat az adatokat, amivel a megfelelő jelentést meg lehet találni és a legmegfelelőbb fordítási ekvivalenst ki lehet választani.

Lássunk egy példát magyar–angol kétnyelvű szótárból. A *Szótár.net* online szótár – előfizetés hiányában – egy egyszerűsített szótári adatbázist kínál fel, mely a fizetős szótártól minőségileg nagyon eltérő szolgáltatást nyújt, de az ingyenes kétnyelvű magyar–angol/angol–magyar szótárak közül így is a jobbak közé tartozik (Felvégi, 2013). A magyar *dob* igéhez az ekvivalensek következő listáját kapjuk:

dob *throw, hurl, fling, toss, cast, sling*

Tehát a szótárhasználat eredménye mindössze a potenciális ekvivalensek listája, mindenféle további információ nélkül, amiből a használó valahogyan ki kell, hogy válassza a számára megfelelőt.

A fenti eredménytől a Magay–Országh (MA) szótár, mely ugyanezen a weblapon keresztül előfizetés vásárlása esetén elektronikusan elérhető (vagy nyomtatott kiadványként megvásárolható), sokkal jobban használható adatokat nyújt:

dob[1] *i* (1) *ált (vmit)* throw*, *[nagy erővel]* hurl (2) *szl [vkin túlad]* throw* sy over/overboard, dump sy (3) *[dobókockával]* **te dobsz!** it's your roll/turn

A fenti szócikkben az *ált* rövidítés önmagában is elég információt nyújt ahhoz, hogy a szótárhasználó egy általánosan használható, „alapértelmezett" ekvivalenst kapjon, a [nagy erővel] vezérszó segít egy speciális és ritkább jelentésárnyalat megtalálásában, a [valakin túlad] vezérszóra ugyanez igaz, a *[dobókockával]* címkével ellátott rész pedig egy mondat méretű egységben mutatja meg a *dob* ige használatának egy hétköznapi, releváns esetét.

Ehhez képest a *dob* igére kapott, vezérszavakat és egyéb információkat nélkülöző *Szótár.net* lista valójában sem tanulási, sem fordítási célra nem alkalmas a magyar anyanyelvű szótárhasználó számára. Bár elvileg nem lehet kifogásunk az ekvivalensek bőséges látványa kapcsán, de a *hurl*, *fling* és a *sling* szavak használata a *dob* ige fordításaként nem feltétlenül és nem olyan eséllyel lesz jó megoldás, mint a *throw* szóé.

Keressük meg az ingyenesen megkapott potenciális fordítási ekvivalensek gyakoriságát a GloWbE korpuszban ~_v* templátumú keresőkifejezéssel (pl. *throw_v**, *hurl_v**; így csak igéket keresünk):

- throw: 78403
- hurl: 1559
- fling: 1314
- toss: 7050
- cast: 54861
- sling: 713

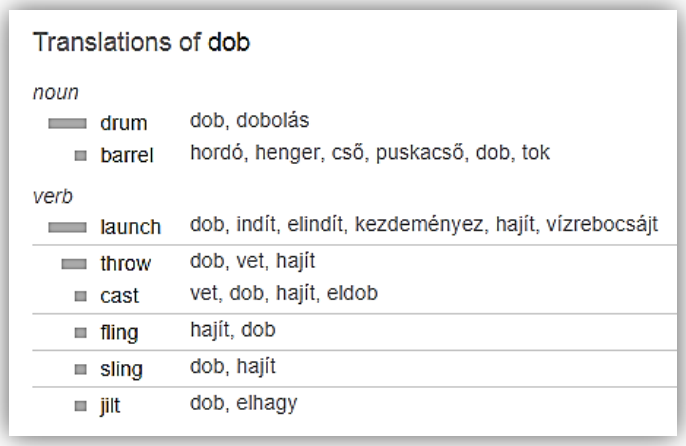
Amennyiben ezeket a gyakorisági adatokat az ingyenes szótárban megkapott ekvivalensek sorrendbe rendezéséhez felhasználjuk, valamint megfelelő

frekvenciasávokat vezetünk be és azokat például csillagozással jelöljük, akkor automatizálható módon ilyen eredményt kaphatunk:

dob *throw***,cast***, toss*, fling, hurl, sling*

Azt nem állíthatjuk ugyan, hogy az így kapott eredmények versenyre kelnének a fizetős szótár professzionális szócikkével, de mindenképpen hasznosabb mind nyelvtanulási, mind fordítási célra az eredeti, ingyenes *Szótar.net* eredménynél.

Az ingyenes megoldások kapcsán nézzük meg azt is, hogy a *Google Fordító* hogyan reagál arra, ha a felhasználó a fordítási mezőbe csak egy szót gépel be (ahogyan azt olyan sok nyelvtanuló szemünk láttára naponta megteszi). A szótározási helyzetet felismerve a *Google Fordító* nem csak egyetlen „fordítást”, hanem az 5. ábrán látható adatsort is visszaadja.



5. ábra: „Szótár” funkció a *Google Fordítás*ban, *dob* szóra keresve, előzetesen magyar → angol fordítási irányt beállítva (http://translate.google.com)

A *Google Fordító* a magyar *dob* szóra 2 főnévi és 6 igei fordítási ekvivalenst adott ebben a példában, és ezek mindegyikének szógyakorisági adatát diagramon szemléltette. Ezen túlmenően mindegyik javaslatát „visszakereste” és megadta azok gyakori magyar megfelelőit. Mivel ez a megoldás nem szótári adatbázissal dolgozik, hanem párhuzamos korpuszokból nyer ki valószínűségi adatokat, így az eredmény itt is elnagyolt, de kiindulási alapnak tulajdonképpen hasonlóan jó, mint amit az *ingyenes* kétnyelvű szótárban jelenleg kap a felhasználó.

A *Google Fordító* néhány összetett szót (pl. *water meter* ’vízóra’) és vonzatos igét (pl. *back up* ’támogat’, stb.) is tud ezzel a módszerrel kezelni. Gépelés közben szinonimákat is keres; angol, francia, spanyol stb. forrásnyelv esetén a beírt szóhoz értelmezést kapunk szótári adatbázisból. Ezen kívül az általa aktuálisan legjobbnak vélt fordítást folyamatosan kijelzi, ahogyan azt egy statisztikai gépi fordítórendszertől (SMT) várjuk. Prószéky (2011) a

webforditas.hu tapasztalatai, felhasználói visszajelzései alapján megjegyzi, hogy sokan talán éppen ezért használják a statisztika gépi fordítást szótár helyett: így nem kell megküzdeniük az ekvivalensek seregével. A webforditas.hu-n a fejlesztők egyébként „megdolgoztatják" (igényesebb eljárásra ösztönzik) ezeket a felhasználókat: egyetlen szóra vonatkozó SMT fordítási kérésre a webhely szótári üzemmódra vált.

5. A szavak együttes előfordulási gyakoriságának mérése

5.1 Elsőrendű együttes előfordulási gyakoriság

A szótárszerkesztők nem kizárólag a különálló szavak („minimális szabad formák"), hanem szókapcsolatok dokumentálásával is foglalkoznak. Ezeket többé-kevésbé kötött, „szinte előre gyártott elemekként" használhatjuk, és együttes előfordulásuk önmagában is „képzeletkeltő" funkcióval bír (Bárdosi, 2012: 7). A kötöttség mértéke különböző lehet, korpuszadatok segítségével pedig mérhető is.

Az éppen jellemzett címszó kollokációi sokszor szabad szemmel is láthatóvá válnak a szó konkordanciáinak különböző kulcs szerint rendezett változataiban. A szemrevételezés azonban nem túl hatékony, és könnyen elmehetünk érdekes jelenségek mellett.

A szavak együttes előfordulási gyakoriságának mérése esetén meghatározzuk a keresett szóalakok vagy lemmák helyzetét a korpuszban, és megszámoljuk azokat az eseteket, amikor két vagy több keresett szó egymás adott méretű környezetében együtt fordul elő. Gyakori az a helyzet, amikor csak az egyik szóról rendelkezünk információval, a másik szóról pedig csak a szófaját vagy akár semmit sem tudunk.

Korpusznyelvészeti értelemben a kollokációkat egy nagyon tág kategóriaként látjuk. A lexikográfus számára azok a kollokációk fontosak, amelyek a szókincs részeként kezelendők, és nem magyarázhatók kizárólag a mondattan saját folyamataival. Ezzel kizárjuk például az egyszerű névelő + főnév kapcsolatot (pl. *the boy*), ugyanakkor nem zárjuk ki a *local/city/country boy* szerkezeteket, melyek 'helyi srác' értelemben a szótárakban is helyet kaphatnak (lásd pl. OLD online). A szótárszerkesztőt számítógépes módszerekkel, kvantitatív vizsgálattal úgy tudjuk segíteni, hogy a potenciálisan figyelemre méltó együttes előfordulásokat listázzuk és a kollokáció korpuszban megfigyelt mért erőssége szerint rendezzük.

A kollokáció erősségének kiszámítása során figyelembe kell vennünk azt, hogy az abban előforduló egyes szavak gyakorisága közvetlenül befolyásolja az együttes előfordulásuk gyakoriságát is. Nagyobb valószínűséggel fog két gyakori szó egymás környezetében előfordulni, mivel egyébként is gyakrabban találkozunk velük. Nekünk azokat az eseteket kell azonosítanunk, amelyek nem csupán ezzel a jelenséggel magyarázhatók, tehát a várt (statisztikai értelemben a

véletlennek köszönhető) együttes előfordulási gyakoriságuknál több együttes előfordulást figyelünk meg.

A kollokáció mértéke a külön-külön fellelt és egymás környezetében megfigyelt előfordulások számának függvénye, eredményül egy valós értéket kapunk, kiszámításakor pedig különböző módszereket követhetünk. Manning és Schütze (Manning–Schütze, 1999) részletesen ismerteti a számítógépes nyelvészeti hátteret. Pecina (2009) a kollokáció erősségének mérésére 55 mértéket mutat be (képletekkel, forrásmegjelöléssel). Ezek némelyike (pl. t-együttható, MI-együttható) gyakran implementált mérték, néhány korpusz webes keresőfelülete közvetlenül is támogatja őket.

A 6. ábrán láthatjuk, hogy az MNSZ2 a *dob* lemmához (korpuszbeli gyakoriság: 84265) milyen lemmákat ad vissza potenciális kollokátumként, ha a keresőszó jobb oldali 3 szavas környezetét vizsgáljuk. A táblázatban az első 12 jelöltet látjuk (a webes felületen ennél sokkal több szerepel), minden sorban az adott lemmára és a *dob* lemmára vonatkozó együttes előfordulási gyakorisági értékkel (Freq), valamint a kiszámított t-score, MI-score és logDice együtthatókkal. A lista rendezési szempontja itt a logDice érték, ami a weblapon változtatható, így különböző módszerekkel kapott kollokáció-jelölteket is figyelembe tudunk venni és össze tudjuk őket hasonlítani.

	Freq	T-score	MI	logDice
piac	5635	74.859	8.500	9.114
ütőhangszer	857	29.269	12.418	8.338
basszus	380	19.470	9.674	7.085
kosár	348	18.601	8.438	6.829
kuka	300	17.291	9.184	6.731
szemét	308	17.464	7.679	6.542
Kata	312	17.565	7.487	6.511
gól	761	27.117	5.878	6.420
labda	367	18.962	6.614	6.382
like-ot	212	14.555	11.458	6.345
üzit	211	14.518	10.879	6.328
utca	945	30.126	5.644	6.324

6. ábra: Kollokáció-jelöltek a *dob* lemmához az MNSZ2 korpusz webes felületén (http://clara.nytud.hu/mnsz2-dev)

Jost és Carus (2003) egy konkrét lexikográfiai projekt (üzleti angol kollokációs szótár) kapcsán, esettanulmány jelleggel mutatja be, hogy ezek az adatok hogyan dolgozhatók fel.

A kiemelt lexikográfiai projektekben is használt *Sketch Engine* szintén támogat számos mértéket, valamint a szavak közti asszociáció kiszámításánál (a *word sketch*ek létrehozása közben) a nyelvtani relációkkal, azok szódisztribúcióra gyakorolt hatásával is tud kalkulálni.

5.2 Másodrendű együttes előfordulási gyakoriság

A szavak elsőrendű együttes előfordulási gyakoriságának a mérése során azt vizsgáltuk, hogy kiválasztott szavak milyen más szavakkal fordultak elő. Ez a (tág értelemben vett) kollokációs jelenségeknek, szintagmatikus viszonyoknak a felderítésében nyújt segítséget. A szavak másodrendű együttes előfordulása („second-order word co-occurrence") az a jelenség, amikor két vagy több szó ugyanazokkal a szavakkal fordul elő (ami egy paradigmatikus kapcsolat), függetlenül attól, hogy egymás környezetében megjelennek-e. Gyakorlatilag tehát azt vizsgáljuk, hogy bizonyos szavak tipikus környezete mennyire hasonló. Ennek a hasonlóságnak az oka a *disztribúciós hipotézis* értelmében jelentéstani eredetű; a szavak ilyen irányú vizsgálatával pedig a disztribúciós szemantika („distributional semantics") foglalkozik.

A másodrendű együttes előfordulás mérése az eredetileg elterjedt eljárás szerint[2] a következő lépésekkel történik:

1. kiválasztjuk azokat a szavakat, amelyeket disztribúciós szempontból össze fogunk hasonlítani (célszavak), és azokat a szavakat, amelyekkel a célszavak tipikus környezetét jellemezni szeretnénk (környezetszavak; ezek száma általában igen magas)
2. a célszavakat nagy korpuszban megkeressük
3. minden egyes találat esetén megvizsgáljuk, hogy a célszó adott környezetében mely környezetszavak fordulnak elő, és a célszó-környezetszó együttes (elsőrendű) előfordulási adatokból a célszavakat jellemző tulajdonságvektorokat készítünk
4. a tulajdonságvektorokban tárolt értékeket súlyozzuk (t-score, MI-score, Dice, stb.) ahhoz hasonlóan, ahogyan a kollokációjelöltekkel eljártunk fentebb
5. a tulajdonságvektorokat ezután összehasonlíthatjuk (pl. vektortávolság, hajlásszög mérésével), klaszterezhetjük, stb.

Pszicholingvisztikai szempontból érdekes, hogy a szavak korpuszadatok alapján kiszámolt hasonlósága korrelál a humán szóhasonlósági kísérletekben kérdőíves módszerrel mért szubjektív szóhasonlósági eredményekkel; magyar adatokért lásd (Tóth, 2013).

A *Sketch Engine* egy kiválasztott célszóhoz a fenti módszerrel automatikusan tud generálni a lexikográfus számára tezauruszt (https://www.sketchengine.co.uk/user-guide/user-manual/thesaurus/). Ez nem olyan adat, ami automatikusan a szótárba kerülhet, csupán egy olyan lista, amely disztribúciós hasonlóság alapján potenciális szinonimákat tartalmaz. Nem kizárólag a tezauruszok, hanem általános szótárak is tartalmazhatnak szinonimákat (pl. MW online, MAC online).

A szinonimák megkeresésén kívül a disztribúciós hasonlósági adatokat potenciálisan felhasználhatjuk bármilyen olyan helyzetben, amelyben a szavak hasonlóságának mérése releváns.

1. Ilyen például a címszavak hasonlóságának mérése. Kilgarriff (1998) alapján tudjuk, hogy az LDOCE3 szótárban a hasonló szavak értelmezésének konzisztenciáját biztosítani kellet, és a 13 részfeladat közül nehézségi sorrendben ez a 7. volt. Ennek a részfeladatnak az első lépését, a hasonló címszavak megkeresését, egy disztribúciós szemantikai eszköz felgyorsítja.
2. Az értelmezésekben használt szavak és a címszavak disztribúciós kapcsolatának mérésére: kontrollált definíciós szójegyzék (CDV) használata esetén a szójegyzékből kiválaszthatjuk azokat a szavakat, amelyek a címszóhoz disztribúciós szempontból hasonlóak. Ez jelenleg egy járatlan terület. Kilgarriff (1998) nehézségi listáján az első helyen szerepel az értelmezések megírása, így ez a lehetőség a későbbiek során figyelmet érdemel.
3. A disztribúciós szemantika egyik területe a többszavas kifejezések kompozicionalitásának mérése (pl. Biemann–Giesbrecht, 2011; Katz–Giesbrecht, 2006; Zhang et al., 2006). A szókapcsolatok kompozicionalitása nagyban befolyásolja azt, hogy érdemes-e (kell-e) őket a szótárban szerepeltetni. Kilgarriff (1998) felmérésének 4. legnehezebb feladata a többszavas kifejezések beválasztása. Ennek a részfeladatnak a megoldásában segíthet egy ilyen új eszköz a lexikográfus „elektronikus munkaasztalán”.
4. Elektronikus szótárakban történő keresés során a szóhasonlóságot is figyelembe tudjuk venni. Erre példa a *OneLook online* metaszótár egyik funkciója (a „kettőspont operátor”). A *:beautiful* keresőkifejezésre a 7. ábrán látható szóhasonlósági listát adja vissza a szótár (a szavak hiperlinkek, így egyszerűen kikereshetők a szótárból). Mint látjuk, a disztribúciós hasonlóság kiszámításával kapott lista (a szótárakban megszokottaktól eltérően) heterogén. A OneLook online a lista további feldolgozásához annyi automatizált segítséget tud csak nyújtani, hogy a találatlistáját szófaj szerint osztályozni tudja (így például a *beautiful* szóhoz kérhetjük az első 100 leghasonlóbb melléknév listázását).

7. ábra: A *:beautiful* keresőkifejezés eredménylistája (1-60. tétel) a OneLook szótárban (http://onelook.com)

6. A szótárhasználó perspektívája: kérdőíves adatok

Az aktív szótárhasználók, valamint a szótárhasználatot (adott esetben szótárdidaktikát) tanító oktatók szempontjából azért fontos ismerni a gyakorisági adatok szótárakra gyakorolt hatását, mert ez hatékonyabb szótárhasználatot eredményez, felfedezhetünk „rejtett” információkat, szükség esetén módosíthatjuk szótárhasználati stratégiáinkat és a szótárak közötti választáshoz is új szempontokat kapunk.

Egy 38 fős, angol szakos egyetemi hallgatókkal (többségükben leendő tanárokkal) végzett anonim kérdőíves vizsgálat során a következőket tapasztaltam:

1. Kedvenc *egynyelvű* szótárként kivétel nélkül ingyenes elektronikus szótárat jelöltek meg (főleg OLD online, LDOCE online és CAD online). Kedvenc *kétnyelvű* szótárakból többfélét említettek, ezek egyike sem volt domináns helyzetben. Többen is valamelyik Magay–Országh nyomtatott szótárat nevezték meg első választásként. Két válaszadó a *Google Fordítót* jelölte meg kedvenc kétnyelvű szótáraként. A kedvenc szótárra vonatkozó kérdés célja az volt, hogy a későbbi kérdésekben egy azonosítható és általuk leginkább ismert szótárra lehessen rákérdezni.

2. A következő kérdés azt mérte, hogy a válaszadók mennyire vannak tudatában annak, hogy kedvenc egynyelvű szótáruk címszólistájának összeállításakor vizsgáltak-e gyakorisági adatokat. A válaszokat 5-fokozatú Likert-skálán gyűjtöttem, a következő eredménnyel.

biztosan igen	*azt gondolja, hogy igen*	*nem tudja*	*azt gondolja, hogy nem*	*biztosan nem*
11%	64%	14%	11%	0%

3. Egy, az előzőhöz hasonló kérdés arról szólt, hogy a hallgatók mennyire vannak tudatában annak, hogy kedvenc kétnyelvű szótáruk címszólistájának összeállításakor alkalmaztak-e gyakorisági adatokat.

biztosan igen	*azt gondolja, hogy igen*	*nem tudja*	*azt gondolja, hogy nem*	*biztosan nem*
6%	67%	18%	6%	3%

   Válaszuk helyességét sok esetben nem tudtam megítélni a megfelelő adatok hiányában, de előfeltételezéseik így is látszanak. Számomra egyébként meglepő volt, hogy ebben az esetben a válaszok aránya nagyon hasonlított az előző kérdésben kapotthoz, miközben a kétnyelvű szótárak között több volt a homályos hátterű, hiányosan dokumentált webszótári projekt.
4. Megkérdeztem őket arról is, hogy az általuk kedvencként megjelölt egynyelvű szótár tudomásuk szerint az értelmezések megírásakor egyszerűsített nyelvezetet használ-e, melyben gyakori szavakat találunk (CDV-k használata, lásd 4.1. fejezet).

inkább igen	*nem tudja*	*inkább nem*
72%	6%	22%

   A következő, ehhez kapcsolódó kérdésben arra válaszoltak, hogy a nyelvtanulók számára egy ilyen megoldás hasznos-e, illetve hasznos lenne-e.

inkább igen	*nem tudja*	*inkább nem*
62%	14%	24%

   A szótárszerkesztési gyakorlatban – éppen az egynyelvű angol tanulói szótáraknál – ez a megoldás teljesen elfogadott és mára igen elterjedtnek számít. A kérdőív alapján ezzel a helyzettel a hallgatók egy része nem volt tisztában, sőt ennek a megoldásnak a megítélése is meglepően vegyes volt.
5. Kedvenc kétnyelvű szótárukkal kapcsolatos kérdés volt, hogy a gyakoribb fordítási ekvivalensek tudomásuk szerint előbbre vannak-e sorolva a ritkábbaknál (lásd 4.3. fejezet).

inkább igen	*nem tudja*	*inkább nem*
89%	0%	11%

Az igenek száma magas, és sokszor hibás feltételezésen alapul.
Szerintük egy ilyen rendezési sorrend hasznos-e?

inkább igen	*nem tudja*	*inkább nem*
76%	15%	9%

6. Azzal kapcsolatban, hogy a jelentések és jelentésárnyalatok sorrendezésénél a szótárszerkesztők szerintük használtak-e gyakorisági információt, a következő adatokat kaptam.

inkább igen	*nem tudja*	*inkább nem*
79%	13%	8%

Szerintük egy ilyen rendezési sorrend hasznos-e?

inkább igen	*nem tudja*	*inkább nem*
80%	20%	0%

A 80%-os támogatottság meggyőző annak ellenére, hogy a 4.2. fejezetben látottaknak megfelelően ebben a kérdésen a lexikográfusok egyébként megosztottak. Egy erre vonatkozó szótárhasználói igény mindenesetre a felmérésből kirajzolódott.

A kérdőíves adatokat egy később elindított szeminárium megtervezéséhez gyűjtöttem, egyrészt a szótárhasználati szokások feltérképezésére, másrészt annak felmérésére, hogy a leendő hallgatók milyen előzetes információkkal rendelkeznek a szótárszerkesztésről. 22 kérdés szerepelt a kérdőíven, ebben a cikkben csak a gyűjtött adatok kis része látható.

A kedvenc kétnyelvű szótár majdnem mindig azon „elavult, pontatlan” szólistára épülő ingyenes online szótárak egyike volt (Felvégi, 2013: 92), melyek szerkesztési elveiről általában keveset tudunk. Sok esetben tulajdonképpen nem is érdemes azt feszegetni, a készítőik vajon milyen lexikográfiai módszerekkel dolgoztak; az angol egynyelvű és az angol–magyar/magyar–angol kétnyelvű *ingyenes* szótárak között mély a minőségbeli szakadék. Ezért arra számítottam, hogy egyrészt a kétnyelvű szótáraknál több lesz a „nem tudja” válasz, másrészt az egynyelvű szótárakról kapott visszajelzés pontosabb, megbízhatóbb lesz. Egyik feltevés sem igazolódott be, mivel az egynyelvű szótárakra vonatkozó ismereteik is hasonlóan hiányosak voltak.

Ezzel kapcsolatban a kiadók felelősségét is meg kell említenünk: az elektronikus szótárakhoz kevés kivételtől eltekintve nem készülnek alapos dokumentációk, ahogyan azt Dringó-Horváth (2011: 143) is megállapítja. Egy-egy jól megírt szótári bevezető beszámol a fontosnak ítélt szerkesztési elvekről, és segít a szótárhasználónak eligazodni a szótárban. Az online szótárakban sokszor keresve sem találunk leírást, dokumentációt, még akkor sem, ha a szótár nyomtatott változatához jól kidolgozott anyagok készültek.

7. Befejezés

A szótárkészítés idő- és munkaigényes folyamat, amit a digitális számítógépes technológiák fejlődése hatékonyabbá és gyorsabbá tett.

Egy – általában hosszú ideig tartó és rengeteg manuális munkát igénylő – lexikográfiai projekt indítása, megtervezése alapvetően befolyásolja a teljes szótárszerkesztési folyamat jellegét és kimenetelét, az elkészült mű tulajdonságait, és meghatározza az adatbázissal később végezhető műveleteket.

A Collins kiadó COBUILD1 szótárának esetében egy teljes lexikográfiai projektet terveztek meg egy korábban ki nem próbált platformon, új megoldásokkal, köztük kvantitatív elemzési módszerekkel, az akkor elérhető legkorszerűbb eszközökkel. Az LDOCE esetében láttuk, hogy a szógyakoriság-alapú „Longman definíciós szójegyzék” (LDV) piacformáló fejlesztés volt. Az Oxford szótársorozat története mutatja, hogy időről-időre szükséges a megújulás, sokszor éppen az új technológiai lehetőségek kihasználása miatt. Szótári adattáruk számítógépes adatbázissá alakítása (az OED2 szerkesztési munkálatai kapcsán) nagy és költséges lépés volt, amit nem lehetett megkerülni. Ugyanez a kiadó az OALD sorozatban saját „kontrollált definíciós szójegyzéket” fejlesztett, szógyakorisági alapra helyezve a tanulói szótárukban található értelmezések nyelvezetét (OALD5), egyben igazodva a Longman által aktívan alakított piaci környezethez.

A szótárak leggyakrabban kereskedelmi termékként készülnek. Minden szót teljeskörűen dokumentálni elvi okokból sem lehet, a gyakorlatban pedig meghatározóak a kiadó által szabott feltételek. Amennyiben csökkenteni kell a feladat mennyiségét, akkor részben a gyakorisági adatokban kell keresnünk a megoldást: ha a ritka szavakat nem választjuk be címszóként, ha a ritka szókapcsolatok értelmezésével nem foglalkozunk, akkor a munka időigényét úgy csökkentjük, hogy a legkevésbé szembeötlő helyen kötünk kompromisszumot. Mindez hasonlít arra, amit az informatika a hang- és videóanyagok tömörítésekor tesz: a percepció szempontjából kevésbé lényeges információt hagyja el, és így csökkenti a kezelendő adatok mennyiségét. A szavak és azok jelentése kapcsán az észlelhetőség fontos kritériuma az adott adat gyakorisága.

Ahogyan az a 6. fejezetből kiderült, a cikk megírásának egyik apropója egy egyetemi szeminárium bevezetése volt. Ezzel kapcsolatos a cikk végére hagyott megjegyzésem. Látjuk, hogy a szótár használói a lehető legkevesebb időt akarják eltölteni szótárhasználattal. Ez számukra egy részfeladat, ami egy másik feladat megoldásához szükséges, kényszerű lépés. Mégis: a kapkodó, felületes munka éles kontrasztban áll azzal, hogy a szótárszerkesztők mennyi és milyen színvonalú munkát fektetnek be egyetlen szócikk létrehozásába is, sokszor milyen mesterműveket alkotnak. A tanárok, a szülők és a szótárakat gondozó kiadók feladata és közös érdeke, hogy az igényes szótárhasználatot és annak

tanítását támogassák, a tanuló pedig a megszerzett tudással azért jár jól, mert a szótárhasználatra szánt időt jobban ki tudja használni, biztosabb kézzel választ szótárt, és pontosabban tudja, hogy mit, hol és hogyan kereshet.

A cikkben hivatkozott szótárak

CAD online = *Cambridge Dictionary Online*, https://dictionary.cambridge.org/
CIDE = Procter, P. (ed.) (1995) *Cambridge International Dictionary of English* (First Edition). Cambridge: Cambridge University Press.
COBUILD1 = Sinclair, J. (ed.) (1987) *Collins COBUILD English Language Dictionary* (First Edition). London/Glasgow: Collins.
COBUILD2 = Sinclair, J. (ed.) (1995) *Collins COBUILD English Dictionary* (Second Edition). London/Glasgow: HarperCollins.
Collins online = *Collins Dictionary online.* http://www.collinsdictionary.com
LDOCE online = *Longman Dictionary of Contemporary English online*, http://www.ldoceonline.com/
LDOCE3 = Summers, D. (ed.) (1995) *Longman Dictionary of Contemporary English* (Third Edition). Harlow: Longman.
MA = Magay T. & Országh L. (szerk.) (2012) Magyar–angol szótár + net + e-szótár. Budapest: Akadémiai Kiadó.
MAC online = *Macmillan Dictionary online*, http://www.macmillandictionary.com
MED1 = Rundell, M. (ed.) (2002) *Macmillan English Dictionary for Advanced Learners* (First edition). Oxford: Macmillan Education.
MW online = *Merriam–Webster Dictionary online*, http://www.merriam-webster.com
NMED = West, M.P. & Endicott, J.G. (eds.) (1935) The New Method English Dictionary. London: Longmans, Green.
OALD5 = Crowther, J. (ed.) (1995) Oxford Advanced Learner's Dictionary of Current English (Fifth Edition). Oxford: Oxford University Press.
OED2 = Simpson, J. & Weiner, E. (1989) *Oxford English Dictionary* (Second Edition). Oxford: Clarendon Press.
OLD online = *Oxford Learner's Dictionaries online*, http://www.oxfordlearnersdictionaries.com/
OneLook = *OneLook Dictionary Search*, http://onelook.com/
Szótár.net = Akadémiai Kiadó *szótár.net* oldala, http://www.szotar.net/
Webfordítás szótár = MorphoLogic *Webfordítás* webszótár, http://www.webforditas.hu/szotar

Felhasznált korpuszok

Brown = Francis, W. N. & Kučera, H. (compilers) (1964) A Standard Corpus of Present-Day Edited American English, for use with Digital Computers. Rhode Island, Providence: Brown University.
GloWbE = Corpus of Global Web-Based English. http://corpus.byu.edu/glowbe/
MNSZ2 = Oravecz Cs., Váradi T. & Sass B. (2014) The Hungarian Gigaword Corpus. In: *Proceedings of LREC 2014*. http://clara.nytud.hu/mnsz2-dev/
NOW = News on the Web korpusz. http://corpus.byu.edu/now/

Irodalom

Bárdosi V. (2012) *Magyar szólások, közmondások adatbázisa*. Budapest: Tinta Könyvkiadó.
Baroni, M., Dinu, G. & Kruszewski, G. (2014) Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. 238–247.
Biemann, Ch. & Giesbrecht, E. (2011) Distributional semantics and compositionality 2011: Shared task description and results. In: *Proceedings of the Workshop on Distributional Semantics and Compositionality (DiSCo)*.
Clear, J. (1987) Computing. In: Sinclair, J. (ed.) *Looking up: An account of the COBUILD project in lexical computing*. London: HarperCollins Publishers. 41–61.

Cruse, A. (2011) *Meaning in Language: An Introduction to Semantics and Pragmatics.* Oxford: Oxford University Press.
Dringó-Horváth I. (2011) Hogyan válasszunk elektronikus szótárat a nyelvtanuláshoz? *Iskolakultúra*. 2011/6–7. 141-156.
Felvégi Zs. M. (2013) Az ingyenes angol–magyar és magyar–angol online szótárakról. *Modern Nyelvoktatás*. 19/4. 80–93.
Howard-Hill, T. H. (1979) *Literary Concordances: A Complete Handbook for the Preparation of Manual and Computer Concordances.* Oxford; New York; Toronto; Paris; Sydney; Frankfurt: Pergamon Press.
Jost, D. & Carus, W. (2003) Computing Business Multiwords: Computational Linguistics in Support of Lexicography. *Dictionaries: Journal of the Dictionary Society of North America*. 24. 59–83.
Katz, G. & Giesbrecht, E. (2006) Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In: *Proceedings of the Workshop on Multiword Expressions*. Sydney: COLING ACL. 12–19.
Kelemen J. (1966) Gépi adatgyűjtés és adatfeldolgozás a lexikográfia szolgálatában. In: Országh L. (szerk.) *Szótártani tanulmányok*. Budapest: Tankönyvkiadó.
Kilgarriff, A. (1998) The hard parts of lexicography. *International Journal of Lexicography*, 11. pp. 51–54.
Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V. (2014) The Sketch Engine: ten years on. *Lexicography. Springer Berlin Heidelberg.* 1 (1), pp. 7–36.
Koplenig, A. & Müller-Spitzer, C. (2014) General issues of online dictionary use. In: Müller-Spitzer, C. (ed.) *Using Online Dictionaries* (Lexicographica Series Maior 145.) Berlin; Boston: Walter de Gruyter. 125–141.
Krishnamurthy, R. (1987) The Process of Compilation. In: Sinclair, J. (ed.) *Looking up: An account of the COBUILD project in lexical computing*. London: HarperCollins Publishers. 62–85.
Lew, R. (2013) Identifying, ordering and defining senses. In: Jackson, H. (ed.) *The Bloomsbury Companion To Lexicography*. London: Bloomsbury Academic.
Manning, C. & Schütze, H. (1999) Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press.
Moon, R. (1987) The Analysis of Meaning. In: Sinclair, J. (ed.) *Looking up: An account of the COBUILD project in lexical computing*. London: HarperCollins Publishers. 86–103.
Morton, H. C. (1994) *The Story of Webster's Third*. Cambridge, New York, Melbourne: Cambridge University Press.
Novák A. (2003) Milyen a jó Humor? In *I. Magyar Számítógépes Nyelvészeti Konferencia*, 138–144. Szeged: SZTE.
Országh L. (szerk.) (1966) *Szótártani tanulmányok*. Budapest: Tankönyvkiadó.
Pajzs J. (1990) *Számítógép és lexikográfia*. Budapest: Magyar Tudományos Akadémia Nyelvtudományi Intézete.
Pecina, P. (2009) *Lexical Association Measures: Collocation Extraction*. Prága: ÚFAL.
Prószéky G. (2011) A szótári világ átalakulási tendenciái az internet megjelentével. *Modern Nyelvoktatás* 17/4, 3–13.
Prószéky G. (2013) How „Truly Electronic Dictionaries" of the 21st Century Should Look Like? In: Stickel, G. & Váradi T. (eds.) *Lexical Challenges in a Multilingual Europe*. Frankfurt am Main; New York; Berlin; Bern; Brüsszel; Oxford; Bécs: Peter Lang Academic Publishers. 51–60.
Prószéky G. & Kis B. (1999) A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, ACL '99*, 261–268, Stroudsburg, PA: Association for Computational Linguistics.
Renouf, A. (1987) Corpus Development. In: Sinclair, J. (ed.) *Looking up: An account of the COBUILD project in lexical computing*. London: HarperCollins Publishers. 1–39.

Summers, D. (1996) Computer lexicography – the importance of representativeness in relation to frequency. In: Thomas, J. & Short, M. (eds.) *Using Corpora for Language Research*. London and New York: Longman. 260–266.

Tóth Á. (2013) Az ember, a korpusz és a számítógép: Magyar nyelvű szóhasonlósági mérések humán és disztribúciós szemantikai kísérletben. *Argumentum* 9, 301–310.

Trakultaweekoon, K., Porkaew, P. & Supnithi, T. (2007) LEXiTRON Vocabulary Suggestion System with Recommendation and Vote Mechanism. In: *Proceedings of Conference of SNLP2007*, 43–48.

Trón V., Halácsy P., Rebrus P., Rung A., Vajda, P. & Simon E. (2006) Morphdb.hu: Hungarian lexical database and morphological grammar. In: *Proceedings of LREC 2006*, 1670–1673.

Xu, H. (2012) A Critique of the Controlled Defining Vocabulary in Longman Dictionary of Contemporary English. *Lexikos* 22. 367–381.

Young, I. D. (1965) *A concordance to the poetry of Byron*. Austin, TX: Pemberton Press.

Zhang, Y., Kordoni, V., Villavicencio, A. & Idiart, M. (2006) Automated Multiword Expression Prediction for Grammar Engineering. In: *Proceedings of the Workshop on Multiword Expressions. Sydney: COLING ACL, 44–52*.

Jegyzetek

[1] Azonban a *lookout* szó értelmezését nem jól fogalmazták meg: *look-out* [noun] [countable] „4. a place for a lookout to watch from”. A szót tehát önmagával értelmezik, az aláhúzott szó pedig olyan hiperlink, ami ugyanerre a bejegyzésre mutat, ’hiperkörkörös’ definíciót eredményezve. Összehasonlításként álljon itt az ehhez a jelentéshez tartozó értelmezés a Macmillan szótárból (MAC online): „a place that is suitable for watching whether someone or something is coming, especially in a dangerous situation”.

[2] Ezt a szakirodalom újabban megszámolós, azaz „count” módszerként nevezi. Alternatívája a konnekcionista gépi tanulást használó „predict” módszer, amely egy neurális hálózat rejtett rétegében keletkező szóábrázolásokat (beágyazásokat, „word embeddings”) használ disztribúciós tulajdonságvektorok helyett. A két módszer összehasonlításáért lásd (Baroni, 2014).