10-16 éves tanulók írásbeli szókincsének gyakorisági szótára[1]

Cs. Czachesz Erzsébet[1], Csirik János[2]

[1]Department of Education, University of Szeged, H6722, Szeged, Petőfi sgt. 30-34.
[2]Department of Computer Science, University of Szeged, H6720, Szeged, Árpád tér 2.

Absztrakt. Az elkészített gyakorisági szótár 4., 6., 8., 10. osztályos tanulók fogalmazásai alapján készült. A minta országosan reprezentatív, a számítógépes elemzés 2170 tanulóra terjedt ki. A teljes korpusz mintegy 600.000 szót tartalmaz.

A gyakorisági szótárak sokféle tudomány és alkalmazási terület fontos forrásai. A továbbiakban a jellegzetes fejlődési irányokat és alkalmazási területeket mutatjuk be röviden. Történetük a 19. század második felére nyúlik vissza, azóta - becslések szerint - mostanáig körülbelül húsz nyelvre mintegy félezer gyakorisági szótár készült. Angol, francia és német nyelvterületen átlagosan öt-hat évenként jelenik meg új.

A magyar nyelvre eddig kevés gyakorisági szótár született. Más szótártípusoknak, így például értelmező, történeti-etimológiai, írói életműveket feldolgozó, asszociációs, tájszótáraknak vannak értékes hagyományai.

Általánosabb célú, a köznyelv valamely rétegének szókészletét feldolgozó gyakorisági szótárunk 1941-ben jelent meg. Nemes Zoltán készítette 401 000 újságnyelvi szövegszó alapján. Ennek közvetlen előzménye az 1933-ban, ugyancsak Nemes Zoltán által jegyzett parlamenti nyelvi gyakorisági szótár. Cser János 1939-ben publikálta szótár formájában a gyermekek szókincsével kapcsolatos kutatásainak eredményeit. A mi munkánk a magyar előzmények közül leginkább ehhez kapcsolódik. 1989-ben jelent meg Füredi Mihály és Kelemen József szerkesztésében a szépprózai gyakorisági szótár, néhány évvel azelőtt pedig a jelen kötet jegyzői által publikált újságnyelvi gyakorisági szótár (Cs. Czachesz Erzsébet és Csirik János, 1986).

Nemzetközileg mérföldkőnek számít ezen a területen Gamble (1861) munkája, aki a kínai ideogrammák gyakoriságát vizsgálta, célja a könyvnyomtatás fejlesztése volt. A gyakorisági szótárak készítésének azonban, a nyelvtudományi alapkutatásokon túl, - a számítógépes szövegfeldolgozás megjelenése és tömeges igénnyé válása előtt - általában valamilyen tág értelemben vett oktatási és nevelési célja volt.

Az első és legfontosabb alkalmazási területük a nyelvoktatás. Az idegennyelv-oktatásban és az anyanyelvi nevelésben is az írás, a beszédkészség, a beszédértéskészség, és az olvasási képesség fejlesztésének tervezésekor gyakran

[1] Az előadás szövege a megjelent szótár (Cs. Czachesz és Csirik, 2002) bevezetőjének rövidített és átdolgozott változata.

fordulnak a kutatók és a tananyagkészítők a gyakorisági szótárakhoz, hogy megtudják, hogy az elsajátítandó szavak milyen gyakorisággal fordulnak elő az adott nyelvben vagy rétegnyelvben (például: West, 1935; Zeno és munkatársai, 1995).

A tizenkilencedik század végén és a huszadik század elején már készültek olyan gyakorisági szótárak, amelyeknek elsődleges célja az oktatásban való hasznosíthatóság volt. Knowles (1904) a vakok olvasókönyveinek összeállítását, Kaeding (1897) és Nemes Zoltán (1933, 1941) a hatékony gyorsírási rendszer tanítását és kialakítását kívánták segíteni. Az idegen nyelvek tanításának nagyobb hatékonysága érdekében, a tanítandó szókészlet kiválasztását először az Amerikai Egyesült Államokban végezték gyakorisági szótárakra támaszkodva. Az Eldridge (1911) által készített szójegyzéknek a célja a bevándorolt munkások nyelvtanulásának megkönnyítése volt. Ennek a szótártípusnak továbbfejlesztett változatai az úgynevezett minimumszótárak, amelyek adott nyelvek szókincséből azokat a leggyakoribb szavakat kívánják meghatározni, amelyek a nyelv valamilyen szintű elsajátításához alapvető fontosságúak (például: Horn, 1926; Bakonyi, 1930; Allwood és Wilhelmsen, 1947; Saukonnen és munkatársai, 1979).

Az utóbbi évtizedekben már többnyire reprezentatív nyelvi mintákon, számítógépes módszerekkel készített többfunkciójú gyakorisági szótárak a jellemzőek (például: Francis és Kucera, 1982; Hall és munkatársai, 1984; Peyawari, 1999; Leech és munkatársai, 2001).

Másik, még mindig közvetlenül az oktatással összefüggő alkalmazási terület az úgynevezett olvashatóság vizsgálata.

Feltételezések szerint egy szöveg olvashatóságának, így tanulhatóságának is az egyik kiemelkedően fontos jellemzője, hogy mennyi benne a ritkán használt, így az átlagos olvasó számára nagyobb eséllyel ismeretlen szó. Az olvasandó szöveg érthetőségének "megjósolására" vállalkoznak a kutatók akkor, amikor különböző olvashatósági formulák alkalmazásával kívánják a tanulásra szánt szövegek ilyen szempontú alkalmasságát vizsgálni (lásd erről részletesen: Harris és Hodges, 1995; Dale és Chall, 1987).

Gyakran használnak az olvasás és a gondolkodás folyamatainak kutatói is, kísérleteik tervezéséhez és végrehajtásához forrásként gyakorisági szótárakat. A szófelismerés mechanizmusainak kutatói például a szavak olvasás útján történő felismerése egyik lényeges befolyásolójának a szó gyakoriságát tartják. Eszerint a gyakoribb szavak felismerése általában még kontextus nélkül is rövidebb reakcióidőt igényel, mint a kevésbé gyakoriaké. Ennek az úgynevezett gyakorisági effektusnak fontos szerepe van például a mentális lexikon felépítésének kísérleti és elméleti modelljeiben is (lásd például: Forster és Chambers, 1973; Whaley, 1978).

A SZÖVEGMINTA

A tanulói szövegek egy reprezentatív fogalmazásvizsgálatból származnak. 1998-ban a József Attila Tudományegyetem (ma Szegedi Tudományegyetem) Pedagógia-Pszichológiai Intézete mellett működő MTA Képességkutató Csoport keretében, a Pedagógiai Tanszék korábbi kutatási eredményeire támaszkodva olyan vizsgálatsorozatot kezdtünk, amelynek segítségével azt kívántuk felmérni, hogy a

magyar közoktatás tanulói az ezredvégen milyen színvonalú képességekkel és készségekkel rendelkeznek. Vizsgálatainkhoz bemért, sztenderdizált mérőeszközöket vettük igénybe.

A programban felmértük az írásbeli kommunikatív képességek, így a fogalmazási képesség színvonalát is. A vizsgálatokat a 4., 6., 8., és 10. osztályosok köréből szervezett országos reprezentatív mintákon végeztük. A minták a 4.-8. évfolyamon a településtípusok, a 10. évfolyamon pedig az iskolatípusok szerint reprezentatívak.

A fogalmazásvizsgálatban minden tanuló két különböző időpontban írt a megadott műfajban és témában egy-egy fogalmazást. Mindegyik műfajban megadtuk a címet, az érvelés esetében két címet is, azon belül további választási lehetőség is volt, aszerint, hogy szeret-e a fogalmazásíró iskolába járni. Az egyik műfaja elbeszélés ("Egy érdekes napom"), a másiké érvelés ("Milyen felnőtt szeretnék lenni?"vagy pedig: "Miért (nem) szeretek iskolába járni?"). Mindegyik fogalmazás írásához 45-45 perc állt a tanulók rendelkezésére. A fogalmazásokat két, egymástól független bíráló hat szempontból (tartalom, szerkezet, stílus, helyesírás, külalak, összbenyomás) értékelte, ezeket az eredményeket használtuk fel (Molnár, Vidákovich és Cs. Czachesz, 2001) a fogalmazási képesség fejlődésének elemzéséhez. Összesen 8670 tanulói fogalmazást elemeztünk, amelyekből évfolyamonként, megyénként és iskolatípusonként külön kisorsoltunk 2170-et, az összes dolgozat negyedét, amelyeket számítógépen rögzítettünk. A rögzítés során változatlanul hagytuk az eredeti formákat, így például helyesírási hibákkal együtt gépeltük le a szöveget. Ezek a fogalmazások képezik az írásbeli tanulói nyelvhasználat szókincsvizsgálatának szövegmintáját.

A SZÖVEGFELDOLGOZÁS MÓDJA

A tanulói szövegek további feldolgozása egy másik kutatási téma és kutatóhely keretében folytatódott. A Szegedi Tudományegyetem Informatikai Tanszékcsoportja mellett működő Mesterséges Intelligencia SZTE-MTA-Kutatócsoport és a MorphoLogic Kft. egy IKTA-pályázat támogatásával, az írott szövegek szövegszavai morfológiai elemzésének és szófaji egyértelműsítésének algoritmikus lehetőségeit vizsgálja. A kutatás tágabb kontextusa az a nemzetközi és hazai informatikai és nyelvészeti kutatási irány és törekvés, amelynek hosszabb távú célja a természetes nyelvek (így a magyar is), gépi feldolgozási lehetőségeinek az előkészítése, illetve megteremtése. A gépi feldolgozás szükségességét nem csupán az Internet és használatának világméretű hódító útja, hanem a gépi (géppel segített) fordítás iránti egyre növekvő igény is indokolja.

Az IKTA-pályázat célja: a kutatók által fejlesztett, úgynevezett tanulási algoritmusok segítségével meghatározott és előre megadott szabályok szerint legyen képes egy program a szövegszavak szótári szavakká való átalakítására. Az ebben a projektben használt korpusz (szövegminta) milliós nagyságrendű, a tanulói fogalmazások összesen körülbelül 600 ezer (de külön kezelt) szövegszavából körülbelül 200 ezer szó méretű anyag is része a teljes korpusznak.

A szövegszavak szótári szavakká való átalakítása folyamán első lépésként minden szó megfelelő szófaji és morfológiai címkéket kapott. (Gépi annotáció.) Az annotáció

alapja a MorphoLogic Kft. által kifejlesztett HuMor elnevezésű magyar nyelvi elemző szoftver, valamint az európai nyelvekre kidolgozott, úgynevezett MSD kódrendszer (Alexin és munkatársai, 1999).

A feldolgozás következő lépéseként a projektben közreműködő egyetemi hallgatók rövid szövegkontextus alapján ellenőrizték, javították és kiegészítették a gépileg kapott annotációk helyességét. A szövegszavak relatív tövének és morfológiai elemzésének ellenőrzésekor a referencia — amikor ez lehetséges volt —a Magyar értelmező kéziszótár (Juhász és munkatársai, 1972) volt.

A szövegszavak szótári szavakká való alakításakor a magyar nyelv jellemzői és a nemzetközi kódolási minták alapján a következő szófaji kategóriákat vettük számításba:

Kategória*	Kód	Kategória	Kód
Melléknév és melléknévi igenév	A	Határozószó, igekötő, határozói igenév	R
Kötőszó	C	Névutó	S
Indulatszó, mondatszó	I	Névelő	T
Számnév	M	Ige	V
Főnév	N	Rövidítés	Y
Névmás	P		

**Megjegyezzük, hogy a szófajilag annotált, egyértelműsített korpusz ennél sokkal részletesebb szófaji információkat tartalmaz. Az alkalmazott MSD kódrendszerben lehetőség volt a ragozási, képzői információk tárolására is.*

A gyakorisági szótár elkészítésekor csak az általános szófaji kategóriákat vettük számításba. A tulajdonneveket nem szerepeltetjük a szótárban, de egy külön tulajdonnévi tárban tároltuk, így a további kutatásra rendelkezésre állnak. A tulajdonnevekből képzett melléknevek viszont megtalálhatóak a szótárban.

A teljes feldolgozott korpusz, amelyhez mindenféle további kutatási célból hozzá lehet férni, a következő web-címen található: http//: www. inf. u-szeged.hu/III/iskolascorpus.html.

A SZÓTÁR FELÉPÍTÉSE

Először abc-rendben adjuk meg a teljes korpusz valamennyi előfordulásra eső összes szavát. Ez a teljes lista lehetővé teszi, hogy az olvasó érdeklődésének megfelelően pótlólagos információkhoz juthasson azokban az esetekben, amikor — helytakarékossági okokból— nem teljes egészében közöljük a részmintákat és a szófaji mintákat.

Ezután felsoroljuk a teljes anyag leggyakoribb 1000 szavát, majd szófajonkénti listákat közlünk. A teljes korpusz szófaji listáin általában az első 600 gyakorisággal bezárólag szerepeltetjük a szavakat, kivéve, ha az előfordulások viszonylag alacsonyabb száma miatt a teljes felsorolást adhatjuk meg.

A szótár második részében életkoronkénti sorrendben közöljük a teljes korpuszban követett eljárás szerint a részminták adatait. Először a negyedikesek, majd a hatodikasok, a nyolcadikasok, végül pedig a tizedik osztályos tanulók fogalmazásainak mintájából a leggyakoribb szavakat (az első 500-500 szót) közöljük. Ezután mindegyik életkorban a szófaji gyakoriságokat adjuk meg.

Az utolsó részben az olvasó összefoglaló táblázatokat találhat, amelyben a teljes minta és az életkoronkénti részminták legfontosabb gyakorisági adatait foglaltuk össze.

Bibliográfia

Alexin, Z.; Váradi, T.; Oravecz, Cs.; Prószéky, G.; Csirik, J.; and Gyimóthy, T. (1999): *FGT-A Framework for Generating Rule-based Taggers.* ILP-99 Late-Breaking Papers, Bled, 24-27 June, p. 1-7. http://www.cs.bris.ac.uk/ilp99/

Allwood, S. and Wilhelmsen, I.(1947): *Basic Swedish Word List.* Rock Island.

Bakonyi, H. (1930): *Die gebrauchlisten Wörter der deutschen Sprache.* München.

Cs. Czachesz Erzsébet és Csirik János (1986): *Újságnyelvi gyakorisági szótár.* Szeged, Budapest, Debrecen; Magyar Pszicholingvisztikai Tanulmányok, IV.

Cs. Czachesz Erzsébet és Csirik János (2002): *10-16 éves tanulók írásbeli szókincsének gyakorisági szótára.* Budapest, Books in Print.

Cser János (1939): *A magyar gyermek szókincse. Gyakorisági és korszótár.* Budapest, Magyar Pedagógiai Társaság.

Dale, S. and Chall, J. S. (1987): *Readability revisited.* New York, McGraw-Hill.

Eldridge, R. C. (1911): *Six Thousand Common English Words.* Niagara Falls.

Forster, K. I. and Chambers, S. M. (1973): *Lexical Acces and naming time.* Journal of Verbal Learning and Verbal Behavior, 12, 627-635.

Francis, W. N. and Kucera, H. (1982): *Frequency Analysis of English Usage: Lexicon and Grammar.* Boston, Houghton Miffin.

Füredi Mihály és Kelemen József (1989): *A mai magyar nyelv szépprózai gyakorisági szótára.* Budapest, Akadémiai Kiadó.

Gamble, W. (1861): *Two Lists of Selected Characters Containing all in the Bible and Twenty-Seven Other Books.* Sanghai.

Hall, W. S.; Nagy, W. E. and Linn, R. (1984): *Spoken Words: Effects on Situation and Social Group on Oral Word Usage and Frequency.* Hillsdale, New Jersey, Lawrence Erlbaum.

Harris, T. L. and Hodges, R. E. (1995): *The Literacy Dictionary. The Vocabulary of Reading and Writing.* Newark, International Reading Association.

Horn, E. (1926): *A Basic Writing Vocabulary.* Iowa City.

Juhász J.; Szőke I.; O. Nagy G.; Kovalowsky m. (1972): *Magyar értelmező kéziszótár.* Budapest, Akadémiai Kiadó.

Kaeding, F. W. (1897): *Haufiigkeitswörterbuch der deutschen Sprache.* Berlin, Steglitz.

Knowles, J. (1904): *The London Print System of Reading for the Blind.* London.

Leech, G.; Rayson, P. and Wilson, A. (2001): *Word Frequencies in Written and Spoken English based on British National Corpus.* Harlow, Longman.

Molnár Edit Katalin, Vidákovich Tibor, Cs. Czachesz Erzsébet (2001): *Writing Development: The Role of School-related and Socioeconomic Factors.* Paper presented at the9th European EARLI-Conference, Switzerland, Fribourg.

Nemes Zoltán (1933): *A magyar parlamenti nyelv leggyakoribb szavai.* Szeged, Az Egységes Magyar Gyorsírás Könyvtára, 66.

Nemes Zoltán (1941): *Szóstatisztika egymillió szótagot felölelő újságszövegek alapján.* Szeged, Az Egységes Magyar Gyorsírás Könyvtára, 190.
Peyawari, A. (1999): *The Core Vocabulary of International English: A Corpus Approach.* Bergen, The Humanities Information Technologies Research Programme.
Saukonnen, P. et al. (1979): *A Frequency Dictionary of Finnish.* Perwoo-Helsinki-Juwa.
West, M. (1935): *Definition Vocabulary.* University of Toronto, Department of Educational Research, Canada.
Whaley, C. P. (1978): *Word-nonword Classification Time.* Journal of Verbal Learning and Verbal Behavior, 17, 143-154.
Zeno, S. M.; Ivens, S. H.; Millard, R. T. and Duvvuri, R. (1995): *The Educators Word Frequency Guide.* New York, TouchstoneApplied Science Associates.