# A Toldi szókészletének eredetéről*

**1. Témaválasztás, előzmények és célok.** Az Arany-év minden bizonnyal serkentőleg hat az olyan kutatásokra, amelyek eddig kevéssé vagy egyáltalán nem érvényesített szempontból közelítenek a gazdag életműhöz.

Közismert, vitán felül álló tény, hogy Arany János kiemelkedő nyelvművész volt. Azzal azonban (tudomásom szerint) még nem foglalkoztak, hogy írásaiban milyen arányban használta a különféle eredetű szavakat. Ezt most már – Beke József szótárának (2017) köszönhetően – Arany teljes költői munkásságára nézve meg lehetne vizsgálni. Ennél azonban szerényebb célt tűztem ki magam elé: a Toldi szókészletének etimológiai statisztikai vizsgálatát, Pásztor Emil Toldi-szótárának (ToldiSz., 1986) alapján.

A magyar etimológiai statisztikai kutatások már majdnem százéves múltra tekinthetnek vissza, de az is igaz, hogy csak kevesen és ritkán vállalkoztak ilyen felmérésre. Az effajta vizsgálódás történetét még az ezredforduló idején tartott előadásomban tekintettem át (Horváth 2000), és később (Horváth 2002: 274) már csak egy-két tétellel egészíthettem ki az ott elhangzottakat. Az akkori szemlém felment a tudománytörténeti előzmények felsorolásának és értékelésének kötelessége alól, így csupán két klasszikussá vált tanulmányt említek meg: Tolnai Vilmosét (1924) és Benkő Lorándét (1962). A mostani elemzésem közvetlenebb előzményeiként utalhatok itt azokra a tanulmányokra, amelyekben magam végeztem etimológiai statisztikai felmérést (Horváth 2000, 2002, 2010, 2016; továbbá nyomtatásban közzé nem tett előadások: Horváth 2004, 2012). – A Benkő-tanulmánynak (1962) és első ilyen tárgyú felmérésemnek (Horváth 2000) az eredményeit egyesítve az egyetemi oktatásban használt szintézisben Zsilinszky Éva állított össze diakrón etimológiai statisztikát (2003: 813–815) a magyar szókészlet változásainak áttekintéséhez.

A vizsgálódás közege szerint az etimológiai statisztikának két fajtáját különböztethetjük meg: a szótári alapút és a használati típusút. Az eddigi felméréseim mindegyike az utóbbi típushoz tartozik.

A használati statisztikák hosszú időn át csakis írott szövegek vizsgálatán alapultak. Korpuszukat hol egy-egy szöveg vagy szövegrészlet alkotta, hol pedig több mű kisebb részleteiből állította össze a szövegmintát az elemző. Saját felméréseim között is van példa mindkét korpusztípusra. – A generációk nyelvhasználatával foglalkozó konferencián tartott előadásomban (Horváth 2016) újdonságképpen spontán beszélt nyelvi szövegekben mértem fel az eredetkategóriák arányait.

* Előadásként elhangzott a Magyar Nyelvtudományi Társaság ülésén 2017. október 3-án. Az írott változat elkészítését az OTKA Új magyar etimológiai szótár. Második ütem című, K 124127 számú projektuma támogatta, melynek munkálatai Gerstner Károly vezetésével folynak az MTA Nyelvtudományi Intézetében. Itt köszönöm meg a gondos munkát és a hasznos észrevételeket tanulmányom két lektorának, Korompay Klárának és Gerstner Károlynak.

Mivel az etimológiai statisztikák elkészítése meglehetősen munkaigényes, érthető, hogy az effajta feladatra vállalkozók általában megelégedtek rövidebb szövegek vagy szövegrészletek elemzésével. Ezzel a hagyománnyal elsőként tanítványom, CSISZÁR GÁBOR szakított: ő etimológiai statisztikai tárgyú szakdolgozatát (2002) Kosztolányi első verseskötetének teljes feldolgozásával írta.

Bennem is felvetődött és megerősödött az a gondolat, hogy a kisebb szövegek, szövegrészletek gyakorisági felmérése után érdemes és szükséges egy nagyobb terjedelmű teljes műnek a feldolgozására vállalkozni. Így jött létre – sakknyelvi kutatásaim részeként is – az első magyar sakk-könyv etimológiai statisztikája (HORVÁTH 2010). Igyekeztem úgy elkészíteni, hogy a sakk világában otthonosakénál szélesebb olvasói kör érdeklődésére is számot tarthasson. Önáltatás volna azonban azt hinni, hogy a könyv speciális, szaknyelvi volta nem szűkítette az érdeklődők táborát.

Megérett tehát bennem az elhatározás, hogy az első próbálkozás után olyan teljes mű legyen a következő etimológiai statisztikám tárgya, amelyet nagyon sokan ismernek és szeretnek. A választásom egyik ötletadója az Arany-év megünneplése volt. Másrészt azért döntöttem éppen a Toldi mellett, mert a valószínűleg legismertebb magyar elbeszélő költemény szókészletének etimológiai felmérését olyan témának tartom, amely valóban széles olvasói rétegeket vonzhat, emellett pedig az oktatásban is felhasználható lehet az általános iskolától kezdve egészen a doktoranduszképzésig.

Akadhatnak olyanok, akik talán szívesebben vennék a Toldi szókészletéből való etimológiai mazsolázgatást, most azonban nem ez a célom, hanem az, hogy a mű egészének szóállományában, valamint a szavak gyakoriságára tekintettel is felmérjem az eredetkategóriák arányait. Céljaim közé tartozik persze annak a megállapítása is, milyenek ezek az arányok a hasonló vizsgálatokban feltártakhoz képest.

**2. Elvek, módszerek és feltevések.** Mint említettem, vizsgálatom alapjául a ToldiSz. szolgált. Emiatt statisztikámat akár szótári alapúnak is lehetne nevezni. A dolog azonban nem egészen így áll. A ToldiSz. ugyanis nem más, mint a Toldi teljes szókészlete szótár formájú feldolgozásban, az adatok előfordulási számának gondos feltüntetésével. Munkám során olyan helyzetben érezhettem magam, mintha magának az elbeszélő költeménynek a szövegéből, azaz használati statisztikát készítve dolgoznék. A szótár azonban az adatok összegyűjtésével és csoportosításával megkímélt a „cédulázástól", pontosabban a szótárban létrehozott címszavakhoz saját feladatomként csak magát az etimológiai minősítést kellett hozzáfűznöm. Ez persze azt is jelentette, hogy elfogadtam segítségül azt a számomra előkészítő jellegű munkát, amelyet PÁSZTOR EMIL tanár úr a szótár összeállításával elvégzett, és meg kellett bíznom az ő pontosságában. Sajnos személyesen már nem köszönhetem meg neki azt, hogy művére ilyen mértékben építhettem, így be kell érnem azzal, hogy elemzésem az ő emléke előtti hálás és tiszteletteljes főhajtás is legyen.

PÁSZTOR EMIL szótára a Toldi teljes szókészletét tartalmazza, összesen 2873 önálló szócikkbe rendezve (ToldiSz. 267). Etimológiai statisztikámat magam is úgy emlegetem, mint a Toldi szókészletének egésze alapján készültet. Ezt az egyszerűség kedvéért és annak jelzésére teszem, hogy nem csupán részletek alapján

dolgoztam. Be kell vallanom azonban, hogy statisztikám a szó szoros értelmében mégsem teljes: korábbi felméréseimhez hasonlóan ez sem tartalmazza a tulajdonneveket. A kihagyott tulajdonnevekről itt nem közlök teljes listát, csak példákat említek közülük: *Bence, György, Sámson, Laczfi, Toldi, Toldiné, Toldi Lőrinc; Bimbó* (ökör), *Rigó* (ló); *Hortobágy, Nagyfalu.* Magukon a tulajdonneveken kívül kihagytam képzős származékaikat és a tulajdonnévi elemű összetett szavakat is: így maradt ki a *Buda, Pest* és *Duna* mellett a *budai,* a *pesti* és a *Duna-part.* (Az utóbbiak kihagyása legalábbis vitatható; erre az egyik lektorom, KOROMPAY KLÁRA is felhívja a figyelmemet. Döntésemet nem is a mostani szóeredet-statisztika, hanem a majdani tőeredet-statisztika indokolja; a két felméréstípus szembeállítását l. később.)

A mű szókészletének többi elemét megtartottam, viszont nem követtem mindenben PÁSZTOR EMIL címszóválasztási elveit. – A ToldiSz.-ban a mutató névmások határozóragos alakjai (*abban, ahhoz, attól, azzá, ennek, erről* stb.) önálló címszavak. Én besoroltam őket a megfelelő névmáshoz, természetesen adatszámukkal együtt. – A ToldiSz.-ban önállóként szereplő határozóragos melléknéveket szintén alapszavukhoz rendeltem: így került például a *hosszan* és a *könnyen* a *hosszú,* illetve *könnyű* címszóhoz. – A ToldiSz. indexes címszavainak egy részét szintén egyesítettem. A szótárban kettő-kettő van például a *derék, majd, örökös, világ* címszavakból; ezekből egyet-egyet hoztam létre. – Szintén összevontam az *azután* és *aztán,* a *gyerek* és *gyermek* címszavakat, valamint a *fi, fia, fiú* címszóhármas tagjait. – Egyesítettem a *fog* igét és segédigét, az *ő* és *ők* névmásokat, továbbá a *van* címszóhoz vontam a ToldiSz.-ban önálló *vala, volna, volt* alakokat.

Néhol ellenkező irányban változtattam, azaz a szótárban közös címszóban szereplő elemeket két szó képviselőivé választottam szét. – Így kerültek például az *azzal* adatai részint az *az* mutató névmáshoz, részint az általam önállóvá tett *azzal* 'akkor' határozószóhoz, az *egyszer* adatai pedig egyrészt az *egy* számnévhez, másrészt az önállósított *egyszer* 'valamely időpontban, hirtelen stb.' határozószóhoz. – A ToldiSz.-ban a *tán* és a *talán* is a *tán* címszó képviselője; én különválasztottam őket. – A ToldiSz.-tól eltérve és az EWUng.-ot követve kettéválasztottam a 'vérér', illetve 'patak' jelentésű *ér* főnevet. – Megszüntettem a szókapcsolati *hím farkas* és *hím szarvas* címszavakat: adataikat a *hím,* illetve a *farkas* és a *szarvas* képviselőinek tekintettem.

A változtatásokhoz tartozik az is, hogy néhányszor módosítottam a PÁSZTOR adta szófaji minősítést. Ez elsősorban melléknévi igenév → melléknév irányú átminősítést jelentett: *ápoló, dagadó, sistergő, főzött.* Van azonban másféle példa is: a 'többé' jelentésű *többször* minősítését ragos számnévről határozószóra változtattam.

Magától értetődik, hogy az etimológiai statisztika készítésekor nem vonhatók össze a *van : lesz, sok : több* típusú szuppletív párok tagjai. (Nem vonta össze őket egyébként PÁSZTOR EMIL sem.)

Korábbi gyakorisági vizsgálataimhoz hasonlóan most is készítettem mind állományi, mind előfordulási statisztikát. Az állományi statisztika csak azzal foglalkozik, melyek azok a szavak, amelyek a korpuszban megjelennek; ismétlődésükre, előfordulási számukra nincs tekintettel. Az előfordulási statisztika viszont természetesen fontos tényezőként veszi figyelembe az ismétlődést.

A ToldiSz. (267) összesítése szerint a szótár 2873 címszavát a Toldiban körülbelül 9900 előfordulási adat képviseli. A kihagyások, átsorolások következtében etimológiai statisztikám szóállománya 2806 elemből áll, 9601 előfordulással.

Azok a kutatók, akik használati típusú etimológiai statisztika készítésére vállalkoztak, hol a szövegben közvetlenül megjelenő szavakat minősítették eredetük szerint, hol visszavezették őket a tövükre vagy másfajta előzményükre, és úgy adtak minősítést, hol pedig mindkétféle módszert alkalmazták. (Részletesebben l. főleg HORVÁTH 2000.) Én már első etimológiai statisztikám összeállításakor úgy gondoltam, hogy mind a szóstatisztika, mind a tőstatisztika elkészítése fontos. – Az első magyar sakk-könyv szókincsét vizsgálva (HORVÁTH 2010) bevezettem egy harmadik statisztikafajtát is: ez a lemmastatisztika. A lemma a szónál kissé elvontabb egység, lényegében úgy tekinthetünk rá, mint a lexikográfiában megszokott címszóra. A szóstatisztikában minden képzett szó külön egységnek számít az alapszavához képest. A lemmastatisztikában ellenben egyes képzőfajtákat (az igenévképzőket, a ható, a műveltető és a szenvedő igék képzőit) lemetszünk, és így adunk minősítést a megmaradt szórésznek. Ennek a „címszavasítási" eljárásnak a neve: lemmatizáció. – Lemmastatisztikát nemcsak a sakk-könyvről készítettem, hanem a nemzedékek spontán beszédének vizsgálatakor is (HORVÁTH 2016), sőt ott a lemmastatisztika önmagában, szó- és tőstatisztika nélkül szerepelt.

A mostani előadásomban a rendelkezésemre álló időbeli, illetve terjedelmi kerethez igazodva a Toldi szavainak és lemmáinak eredetstatisztikáját mutatom be. Ez azonban nem jelenti azt, hogy a tőstatisztika összeállításáról és ismertetéséről lemondok, csupán más alkalomra hagyom.

A szavak (és lemmák) eredetminősítéséhez – mint a korábbiakban is – elsősorban az EWUng.-ot hívtam segítségül, és persze szükség esetén figyelembe vettem a szavak morfémaszerkezetét. Az EWUng.-beli minősítésen csak igen ritkán változtattam, például amiatt, hogy az ottani egyetlen *egy* címszó helyébe kettőt vettem fel: a számnévtől és névmástól elválasztottam a határozatlan névelőt, és az utóbbit szófajváltás eredményének tekintettem (részben a mutató névmásnak és a határozott névelőnek az elkülönítéséhez igazodva). – Más példaként megemlítem azt is, hogy tanulmányom kéziratában az EWUng.-ot követve (noha nem szívesen) megtartottam a „hangátvetés"-t mint eredetminősítést; KOROMPAY KLÁRA lektori bírálatának hatására örömmel hagytam el. (Érintettje egyébként itt csak a *sárga* szó volt.) – Az EWUng. után megjelent származtatási ötleteket (nyilvánvalóan vitatható döntéssel, de elvszerűen) „kanonizálatlanságuk" miatt, azaz összefoglaló etimológiai szótár által egyelőre (?) meg nem erősített voltuk miatt nem vettem figyelembe.

Eddigi elemzéseim többségében az egyes statisztikatípusokat két változatban készítettem el. Ezek a bizonytalan és vitatott eredetmagyarázatok kezelésében különböztek egymástól. Az I. változatban a bizonytalan és a vitatott eredetű elemek önálló kategóriákként szerepeltek, hasonlóan például az összetételekhez, az olasz jövevényekhez vagy az ismeretlen eredetűekhez. A II. változatban viszont megszüntettem a bizonytalanok és a vitatottak kategóriáját. A bizonytalan származásúakat az EWUng. szerint esetleg számításba vehető eredeztetésüknek megfelelően láttam el minősítéssel, míg a vitatottak az EWUng.-ban 1. helyen említett

eredetminősítést kapták. Ehhez hasonlóan számoltam fel az alapnyelvi örökségen belüli tisztázatlan rétegnek, a belső keletkezésűek tisztázatlan kialakulásmódú csoportjának, valamint a tisztázatlan átadó nyelvből való jövevényeknek a kategóriáját is. (A finomabb részletekhez l. főleg HORVÁTH 2000: 180, 321–322.) – Az I. és a II. változatot a Toldi etimológiai statisztikájában is megkülönböztetem: mind a szavakra, mind a lemmákra nézve.

A valószínű származtatásokat, akárcsak korábbi elemzéseimben, egyenértékűeknek tekintettem a biztosakkal (ennek megokolásához l. HORVÁTH 2000: 176).

A kategorizálás más részleteiben is lényegében alkalmazkodtam régebbi statisztikáimhoz. Az alapnyelvi örökség rétegeit megkülönböztettem egymástól. Nem differenciáltam a konkrétabb szóátadó nyelvek szerint az iráni, a török és a szláv jövevényeket, az újlatinokat viszont igen. – Az igekötős igéket az összetett szavak közé soroltam. Az *elfeledhet, előrehajolva* típust viszont, amelyben képző is van az igekötős igén, *elfeled* + *-het, előrehajol* + *-va* tagolásúnak, vagyis származéknak, nem pedig összetételnek tekintettem. – A további részletekkel kapcsolatban részint az eddigi írásaim, részint a statisztikák bemutatásakor következő példák nyújtanak felvilágosítást.

Mire számítottam az eredetkategóriák Toldi-beli arányait illetően munkám megkezdése előtt? Eddigi szóstatisztikáimban mindenütt a belső keletkezésű elemek uralkodtak; Arany nyelvhasználatára, valamint a Toldi műfajára, témájára, közegére gondolva azt feltételezhettem, hogy ezúttal a szokottnál is erősebb lesz a dominanciájuk. A jövevényszavak közül inkább a magyarba simábban beilleszkedő, kevésbé idegennek érzett török és szláv elemek jelenlétére gondolhattam, mint a németekére vagy a latinokéra. Biztosra vettem, hogy nemzetközi szó egyáltalán nincs a műben, vándorszó viszont lehet. Valószínűtlennek tartottam a tükörszók megjelenését. Kíváncsian vártam és többesélyűnek tartottam viszont a származékszók (képzett szavak) és az összetett szavak küzdelmét.

**3. A Toldi szavainak eredetstatisztikája.** Az I. típusú, azaz a „bizonytalan" és „vitatott" minősítéseket figyelembe vevő szóeredet-statisztikát az 1. táblázat mutatja be.

**1. táblázat**

A Toldi I. típusú szóeredet-statisztikája

| **Eredet** | **Szó** | **%** | **Előfordulás** | **%** |
|---|---|---|---|---|
| uráli | 43 | 1,53 | 378 | 3,94 |
| finnugor | 73 | 2,60 | 830 | 8,64 |
| ugor | 35 | 1,25 | 159 | 1,66 |
| tisztázatlan rétegből | 18 | 0,64 | 153 | 1,59 |
| kétforrású örökség | 1 | 0,04 | 18 | 0,19 |
| örökség magyar képzővel | 68 | 2,42 | 368 | 3,83 |
| **örökség összesen** | **238** | **8,48** | **1906** | **19,85** |
| származék | 820 | 29,22 | 1269 | 13,22 |

| | | | | |
|---|---|---|---|---|
| fiktív tőből | 63 | 2,25 | 150 | 1,56 |
| képzőcserés | 1 | 0,04 | 1 | 0,01 |
| elvonás | 14 | 0,50 | 22 | 0,23 |
| szórövidülés | 9 | 0,32 | 18 | 0,19 |
| szóösszehúzás | 5 | 0,18 | 13 | 0,14 |
| jel- vagy ragszilárdulás | 220 | 7,84 | 986 | 10,27 |
| önállósulás | 4 | 0,14 | 18 | 0,19 |
| szófajváltás | 57 | 2,03 | 1255 | 13,07 |
| jelentéselkülönülés | 2 | 0,07 | 3 | 0,03 |
| szóhasadás | 13 | 0,46 | 147 | 1,53 |
| ikerszó | 4 | 0,14 | 4 | 0,04 |
| összetétel | 817 | 29,12 | 1641 | 17,09 |
| szóösszevonás | 3 | 0,11 | 11 | 0,11 |
| tükörösszetétel | 1 | 0,04 | 1 | 0,01 |
| onomatopoetikus | 57 | 2,03 | 132 | 1,37 |
| belső, de tisztázatlan | 19 | 0,68 | 203 | 2,11 |
| **belső keletkezésű összesen** | **2109** | **75,16** | **5874** | **61,18** |
| permi | 1 | 0,04 | 6 | 0,06 |
| iráni | 9 | 0,32 | 37 | 0,39 |
| török | 61 | 2,17 | 201 | 2,09 |
| német | 31 | 1,10 | 71 | 0,74 |
| latin | 15 | 0,53 | 28 | 0,29 |
| olasz | 13 | 0,46 | 22 | 0,23 |
| román | 1 | 0,04 | 2 | 0,02 |
| szláv | 98 | 3,49 | 374 | 3,90 |
| tisztázatlan nyelvből | 4 | 0,14 | 8 | 0,08 |
| közvetített | 5 | 0,18 | 8 | 0,08 |
| jövevény magyar képzővel | 7 | 0,25 | 10 | 0,10 |
| vándor | 9 | 0,32 | 23 | 0,24 |
| **jövevény összesen** | **254** | **9,05** | **790** | **8,23** |
| bizonytalan | 93 | 3,31 | 527 | 5,49 |
| vitatott | 34 | 1,21 | 263 | 2,74 |
| ismeretlen | 78 | 2,78 | 241 | 2,51 |
| **tisztázatlan összesen** | **205** | **7,31** | **1031** | **10,74** |
| **Összesen** | **2806** | | **9601** | |

A szóállomány háromnegyed része belső keletkezésű. A többi főkategóriához képest ez kiemelkedő részesedés, más statisztikáimmal összevetve azonban nem rendkívüli: lényegében megfelel a 20. század végi elbeszélésekben mértnek

(vö. HORVÁTH 2000: 317), az első (18. századi) magyar sakk-könyvben találtnál (vö. HORVÁTH 2010: 424) pedig 10 százalékponttal kisebb.

A másik három főkategória részesedése 10% alatti. A jövevények, az alapnyelvi örökség képviselői és a tisztázatlan eredetűek szorosan követik egymást.

A kategóriák összességét tekintve – és így persze a belső keletkezésűek között is – a képzett származékok és az összetett szavak óriási fölénnyel, egymással szinte azonos, a 30%-ot megközelítő részesedéssel vezetik a listát. Rajtuk kívül egyetlen kategória sem éri el a 10%-ot; a harmadik helyezettnek, a megszilárdult ragos és jeles alakulatoknak a részesedése is csak 8%-nyi. – Különösen az összetett szavak nagyarányú jelenléte érdemel figyelmet, mivel a 20. század végi elbeszélésekben 10 százalékpontnyi hátránnyal követik a származékokat.

A jövevényszavak között a várakozásnak megfelelően a szlávoké a vezető szerep; az sem meglepő, hogy a törökök következnek utánuk. – Az alapnyelvi örökség rétegei közül a finnugornak a képviselete a legnagyobb. – Az ismeretlen eredetű szavak részesedése nagyon hasonló a 20. század végi elbeszélésekben mérthez (vö. HORVÁTH 2000: 317).

Mivel a Toldi szóállományának összesen 2806 tagját 9601 előfordulási adat képviseli, egy-egy szóra átlagosan 3,42 előfordulás jut. (Összehasonlításul: a 18. századi sakk-könyvben 4,26 az átlag, a 20. század végi elbeszélésekben pedig 1,92.)

Az ismétlődés a Toldiban (akárcsak a másik két említett korpuszban) az alapnyelvi örökség elemeire a legjellemzőbb: előfordulási hányadosuk 8,01. Utánuk ez a sorrend: tisztázatlan eredetűek: 5,03; jövevények 3,11; belső keletkezésűek: 2,79.

A hányadosokat látva érthető, hogy a szóelőfordulások statisztikája erősen átrendeződik a szóállományéhoz képest.

Az előfordulások között is nagy fölénnyel állnak az élen a belső keletkezésűek, de itteni részesedésük alig haladja meg a 60%-ot. A sakk-könyvnek és a 20. századi elbeszéléseknek az anyagát feldolgozva is azt tapasztalhattam, hogy a belső keletkezésű szavak részesedése az előfordulások között csekélyebb az állománybelinél, de a visszaesés mértéke amazokban kisebb.

Az átlagosnál sokkal sűrűbb ismétlődésüknek köszönhetően az alapnyelvi örökség elemeinek részesedése a szóelőfordulások között 20%-nyi.

A jövevényszavak ismétlődése átlagos mértékű. Előfordulási arányuk is hasonló a szóállománybelihez.

A tisztázatlan eredetűek az átlagosnál többször ismétlődnek, így az előfordulások között a részesedésük 10% fölé emelkedik.

A szóelőfordulásokból négy kategóriának van 10%-nál nagyobb részesedése. Mindegyik a belső keletkezésűek hez tartozik.

A lista élén itt az összetett szavak állnak 17%-os képviselettel. Ismétlődési hányadosuk (2,01) kisebb a korpuszátlagnál; ezzel összefüggésben részesedésük 12 százalékponttal elmarad az állománybelitől. Első helyüket annak köszönhetik, hogy a származékok ismétlődése még ritkább (1,55). – Az összetett szavak gyakorisági listáját olyan elemek vezetik, amelyeknek az összetett volta már nagyon régen elhomályosult; itt és a következőkben zárójelben a Toldi-beli előfordulási számot adom meg: *is* (131), *és* (89), *sem* (54), *én* (28). Utánuk a sorban hozzájuk hasonlóan szintén nem fogalomszói elemek jönnek: *mintha* (28), *azután* (24), *ha-*

*nem* (24), *aki* (22). Az összetett fogalomszók közül a Toldiban az összetételként elhomályosult *ember* (17) a leggyakoribb, a világos fogalomszói összetételek közül pedig az *édesanya* (14). Az igekötős igék leggyakoribbja a *megöl* (11).

A második helyen álló származékokra még kevésbé jellemző az ismétlődés, mint az összetételekre. – Kiemelkedően leggyakoribb reprezentánsuk az *isten* (36), utána a *szól* (14), az *erős* (13) és a *kutya* (13) következik. Nem fogalomszói képviselőikből a *mind* (9) és az *olyan* (8) jelenik meg legtöbbször.

Más etimológiai statisztikákból is jól ismert jelenség, hogy a szóállományihoz képest az előfordulási statisztika leginkább a szófajváltásnak kedvez. Ennek a kategóriának az egy szóegyedre eső ismétlődési hányadosa más korpuszokban is igen nagy; a Toldiban 22,02. Nem csoda, hogy az előfordulások között a szófajváltás a kategóriák listájának a harmadik helyére kerül. – Leggyakoribb képviselői a határozott névelők: *a* (666), *az* (157). Utánuk kötőszók állnak: *hogy* (94), *de* (75). Itt következik az *egy* határozatlan névelő (59), majd megint két kötőszó: *mint* (46), *vagy* (25). Az ide tartozó fogalomszók közül a *farkas* (17) a leggyakoribb, az igenévi származásúakból pedig a *mező* (6).

Az átlagosnál sűrűbb ismétlődésnek (4,48) köszönhetően a megszilárdult ragos és jeles alakulatok kategóriája a szóelőfordulások között átlépi a 10%-os küszöbértéket. – A ragszilárdulásos típusnak a leggyakoribb képviselője a *ha* kötőszó (54), ezt a *most* (40) és a *minden* (26) követi. A jelszilárdulásos típus legtöbbször megjelenő reprezentánsai pedig a *maga* (52) és a *neki* (34).

A belső keletkezésűek körében az eddig bemutatottakon kívül még négy olyan kategória van, amelynek az előfordulási részesedése meghaladja az 1%-ot.

A fiktív tövű származékszavak között abszolút, illetve relatív tőből létrejötteket is találunk, a korpuszátlagnál ritkább ismétlődéssel (2,38). Listájukat a *szörnyű* (14) és a *kegyelem* (10) vezeti.

A szóhasadás reprezentánsainak feltűnően nagy az ismétlődési hányadosuk (11,31). Néhány közülük különösen gyakran fordul elő: *pedig* (43), *mert* (31), *még* (29), *vesz* (20).

A fiktív tövűekhez hasonlóan az onomatopoetikus elemek ismétlődése is kisebb az átlagosnál (2,32). Közülük csak a *jaj* igazán gyakori (21). A hangutánzó igék közül a *kiált* jelenik meg a legtöbbször (5).

Az olyan belső keletkezésű elemeknek az ismétlődési hányadosa (10,68), amelyeknek a keletkezésmódja tisztázatlan, megközelíti a szóhasadás kategóriájáét. Elég sok olyan akad közöttük, amely különösen gyakorinak mondható: *úgy* (47), *így* (35), *egy* számnév és névmás (32), *hogy* ’hogyan’ (24), *már* (24), *majd* (22).

A többi belső keletkezésű szókategória 1%-nál ritkábban fordul elő. Ez persze összefügg azzal, hogy képviselőik között egyetlen olyan sincsen, amely feltűnően sűrűn (legalább 20-szor) bukkan fel a Toldiban. – Az elvonás leggyakoribb képviselője az *arc* (4), a szórövidülésé a *kicsi* (4), a szóösszehúzásé (vagyis az olyan rövidülésé, ahol a szó nem a végén, hanem a belsejében rövidült meg) a *tán* (8), a szóösszevonásé (csonkulásos összetételé) pedig a *midőn* (9). Az önállósulás képviselői közül a *testvér* elég sokszor tűnik fel (14). Jelentéselkülönüléssel a szóállomány két tagja keletkezett: a *fok* (késé, váré) kétszer, az *avas* ’bozótféle’ egyszer fordul elő. A korpusz négy ikerszavának mindegyike csak egyszer-egyszer

jelenik meg: *dibdáb, dínomdánom, hórihorgas, tarkabarka.* A képzőcserének a szóállományban csupán egyetlen képviselője van, egy adattal: a *hervatag.* – Noha előzetesen nem számítottam tükörfordítás megjelenésére, mégis akad egy belőle a Toldiban, ha egyetlen adattal is: ez a latin mintájú *feltétel.*

Az alapnyelvi örökség rétegei közül – akárcsak a szóállományban – az előfordulásokat tekintve is a finnugoré a vezető szerep. Részesedése önmagában is majdnem 9%-nyi. Ismétlődési hányadosa is nagyon tekintélyes (11,37). – Képviselői közül kettőnek a gyakorisága egészen kiemelkedő: a *nem* tagadószóé (181) és a *van* igéé (151). Ráadásul még hét reprezentánsának van 20-nál több adata: *ő* (39), *szép* (39), *ne* tiltószó (32), *fiú* (27), *kéz* (26), *öcs* (23), *ház* (22). A képhez persze az is hozzátartozik, hogy 23 finnugor eredetű elem csupán egy adattal jelenik meg a Toldiban. (Többek között ezért nem még nagyobb az ismétlődési hányados.)

Az uráli származású réteg képviselete az előfordulások között csaknem 4%. Az ismétlődési hányados itt is nagyobb a korpuszátlag duplájánál (8,79). – Az uráli eredetű szóállomány is bővelkedik gyakran használt elemekben; tagjai közül hétnek is 20-nál több adata van: *megy* ige (36), *szem* (36), *fej* főnév (29), *lát* (27), *szív* főnév (27), *tud* (26), *e* névmás (24).

Az ugor réteg képviselete sokkal kisebb, jóllehet ennek az ismétlődési hányadosa is átlagon felüli (4,54). – Reprezentánsai közül kettő igen gyakori: a *jó* (34) és a *szó* (27).

Az olyan alapnyelvi eredetű elemek, amelyekkel kapcsolatban nincs egészen tisztázva az, hogy melyik rétegből öröklődtek, az előfordulások között nagyjából olyan részesedésűek, mint az ugorok; ismétlődési hányadosuk viszont sokkal nagyobb azokénál (8,50), inkább az uráliakéhoz áll közel. – Két kiemelkedően gyakori képviselőjük a finnugor vagy esetleg az uráli korból való örökség: a *ki* névmás (50) és a *két* számnév (36).

A „kétforrású örökség" minősítést a *hall* ige kedvéért kellett felvennem. Ez két alapnyelvi elemnek a sajátos, összetartó fejlődésével jött létre (a részleteket l. az EWUng.-ban). Ez az ige a Toldiban 18-szor bukkan fel.

Az „örökség magyar képzővel" minősítés azt jelenti, hogy a valamely alapnyelvi rétegből örökölt (és fiktív) tőhöz magyar képző járul, viszont a képzett morfémaegyüttes (vagyis szó) egésze nem rekonstruálható az alapnyelvre. Az ilyen szavakat a tő rétegbeli hovatartozása (uráli, finnugor, ugor) szerint nem differenciáltam. Ennek a kategóriának a részesedése a szóállományt tekintve alig marad el a finnugorétól. Az ismétlődési hányadosa nagyobb ugyan a korpuszátlagnál (5,41), de a finnugor és az uráli rétegénél sokkal kisebb, így előfordulási részesedése kissé elmarad az uráli elemekétől. – Az ide tartozó szavak közül három fordul elő a Toldiban 20-nál többször: *lesz* (55), *mond* (34), *anya* (26).

A jövevényszavak között – még inkább, mint a szóállományban – a szláv eredetűeké a vezető szerep. Előfordulási részesedésük megközelíti a 4%-ot; körülbelül akkora, mint az alapnyelvi rétegek közül az uráliaké. Ismétlődési hányadosuk (3,82) valamivel nagyobb a korpuszátlagnál. – Ez az egyetlen olyan jövevényszóréteg, amelynek vannak a Toldiban 20-nál nagyobb adatszámú képviselői. Öt ilyen is akad: *király* (41), *cseh* (26), *vitéz* (26), *dolog* (22), *szolga* (21).

A török elemek részesedése kisebb a szlávokénál, a többi jövevényszórétegét azonban jóval felülmúlja. Ismétlődési hányadosuk (3,30) nagyon közeli a korpuszátlaghoz. – Négy olyan képviselőjük van, amelynek előfordulási száma 10 és 20 közötti: *erő* (14), *kis* (13), *bika* (12), *kar* 'végtag' (11).

A jövevényszavak közül a szóállományra nézve még a németeké nagyobb 1%-nál, de ismétlődésük ritkább az átlagosnál (2,29), így az előfordulások között már nem érik el az 1%-os küszöbértéket. – Leggyakrabban előforduló képviselőik adatszáma is 10 alatti: *marha* (7), *kanna* (6), *pár* (6), *tarsoly* (6).

Az irányi jövevények ismétlődése az átlagosnál sűrűbb (4,11), de kifejezetten gyakori szó nincs közöttük. – A legtöbbször megjelenő képviselőik: *asszony* (10), *kard* (7), *vár* főnév (5).

A Toldiban a latin jövevényszavak ritkán ismétlődnek (1,87). Még a legtöbbször felbukkanó reprezentánsaik sem gyakoriak: *mód* (6), *cifra* (4). – Ugyanezt mondhatom az olasz jövevényekről (1,69): *part* (5), *pajzs* (3). – A korpusz egyetlen román eredetű eleme a *cimbora* (2).

Permi jövevényszó szintén csupán egy van: *kenyér* (6).

Nincs tisztázva, honnan került jövevényként a magyarba a *köntös* és a *rúd* (3-3), továbbá a *bér* és a *tornác* (1-1).

Öt jövevényszó esetében (az EWUng. szerint) feltételezhető, hogy nemcsak egy nyelvből, hanem talán egy másik nyelv közvetítésével is jutott a magyarba: a latinból jövő *alamizsna* és *pogány* (2-2 adat) esetében szláv, az olasz származású *tréfa* (2) átvételében latin, a németből való *bitang*-gal kapcsolatban cseh, a török származású *dandár* esetében pedig szerbhorvát közvetítéssel lehet számolni. – Statisztikámban a talán közvetített jövevények kategóriáját az átadó nyelvek szerint nem differenciáltam.

Egységesen, az átadó nyelvektől függetlenül kezeltem azt a jövevényszótípust is, amelyben az átvétel magyar képző hozzáadásával (de nem szokásos honosító igeképzővel!) történt. – Közülük a Toldiban két-két alkalommal bukkan fel a *borul,* a *kocsmáros* és a *mészáros,* egyszer pedig a *bosszant,* a *gyaláz,* az *óbégat* és a *pallos.*

Tágabb értelemben természetesen a jövevényekhez tartoznak a vándorszók is. Ismétlődésük ritkább az átlagosnál (2,56). – Leggyakoribb képviselőik: *paripa* (5), *levente* (4), *tarisznya* (4).

A tisztázatlan eredetűek közül a bizonytalan származásúak ismétlődési hányadosa (5,67) nagyobb a korpuszátlagnál; ennek köszönhetően a részesedésük az előfordulások között jelentősebb az állománybelinél: 5% feletti. – Két képviselőjük adatszáma kiemelkedően nagy; mindkettő talán finnugor származású: *s* (132), *nagy* (85). Az olyan bizonytalan eredetűek közül, amelyekkel kapcsolatban a belső keletkezés jöhet szóba, a *föld* a leggyakoribb (talán származék; 20 adat); amelyeknek a származtatásában pedig jövevényszó voltuk vetődik fel, a *gyermek* (talán török; 15) és a *nád* (talán iráni; 10).

A vitatottak ismétlődési hányadosa (7,74) a bizonytalan eredetűekénél is nagyobb, de előfordulási számuk így sem éri el amazokénak a felét sem. – Itt is két olyan elem vezeti a gyakorisági listát tekintélyes adatszámmal, amelyek esetében elsősorban a finnugor származás jöhet számításba: az *az* (105) és *ez* (43) mutató

névmások. Az ide tartozó leggyakoribb fogalomszóknak az EWUng.-ban 1. helyen álló eredeztetése viszont a török: *öreg* (20), *úr* (14).

A Toldi ismeretlen eredetű szavai az átlagosnál kissé ritkábban ismétlődnek (3,09), de részesedésük a szóelőfordulásokat tekintve éppen úgy 2 és 3% közötti, mint a szóállományban. – Kiemelkedően gyakori képviselőjük csupán egy van: *csak* (61); ezt a *nap* 'Sonne, Tag' (14) és a *hisz* ige (10) sokkal lemaradva követi.

A Toldi II. típusú szóeredet-statisztikáját – a 2. pontban tárgyalt módon – a bizonytalan és vitatott kategóriák felszámolásával hoztam létre. Az átminősítéssel szintén megszűnt a tisztázatlan rétegű örökségnek, a tisztázatlan kialakulásmódú belső keletkezésűeknek és a tisztázatlan nyelvből átvett jövevényszavaknak a kategóriája; sőt így járt a közvetített jövevényszavaké is, mivel egyik reprezentánsuknak sem biztos a közvetített volta. (Az utóbbiak a II. statisztikában „fő" átadó nyelvük képviselőivé váltak.) – Az átrendezéssel kialakult képet a 2. táblázat mutatja be.

**2. táblázat**

A Toldi II. típusú szóeredet-statisztikája

| **Eredet** | **Szó** | **%** | **Előfordulás** | **%** |
|---|---|---|---|---|
| uráli | 49 | 1,75 | 397 | 4,13 |
| finnugor | 103 | 3,67 | 1365 | 14,22 |
| ugor | 46 | 1,64 | 204 | 2,12 |
| kétforrású örökség | 1 | 0,04 | 18 | 0,19 |
| örökség magyar képzővel | 95 | 3,39 | 450 | 4,69 |
| **örökség összesen** | **294** | **10,48** | **2434** | **25,35** |
| származék | 831 | 29,62 | 1335 | 13,90 |
| fiktív tőből | 83 | 2,96 | 191 | 1,99 |
| becéző | 1 | 0,04 | 14 | 0,15 |
| képzőcserés | 2 | 0,07 | 2 | 0,02 |
| elvonás | 16 | 0,57 | 24 | 0,25 |
| szórövidülés | 11 | 0,39 | 21 | 0,22 |
| szóösszehúzás | 5 | 0,18 | 13 | 0,14 |
| jel- vagy ragszilárdulás | 225 | 8,02 | 1146 | 11,94 |
| önállósulás | 4 | 0,14 | 18 | 0,19 |
| szófajváltás | 64 | 2,28 | 1263 | 13,15 |
| jelentéselkülönülés | 5 | 0,18 | 8 | 0,08 |
| szóhasadás | 15 | 0,53 | 153 | 1,59 |
| ikerszó | 4 | 0,14 | 4 | 0,04 |
| összetétel | 818 | 29,15 | 1648 | 17,16 |
| szóösszevonás | 3 | 0,11 | 11 | 0,11 |
| tükörösszetétel | 1 | 0,04 | 1 | 0,01 |
| onomatopoetikus | 59 | 2,10 | 134 | 1,40 |

| **belső keletkezésű összesen** | **2147** | **76,51** | **5986** | **62,35** |
|---|---|---|---|---|
| permi | 2 | 0,07 | 7 | 0,07 |
| iráni | 15 | 0,53 | 64 | 0,67 |
| kaukázusi | 1 | 0,04 | 1 | 0,01 |
| török | 77 | 2,74 | 280 | 2,92 |
| német | 36 | 1,28 | 86 | 0,90 |
| latin | 18 | 0,64 | 34 | 0,35 |
| francia | 1 | 0,04 | 1 | 0,01 |
| olasz | 15 | 0,53 | 31 | 0,32 |
| román | 1 | 0,04 | 2 | 0,02 |
| szláv | 102 | 3,64 | 389 | 4,05 |
| jövevény magyar képzővel | 10 | 0,36 | 22 | 0,23 |
| vándor | 10 | 0,36 | 24 | 0,25 |
| **jövevény összesen** | **288** | **10,26** | **941** | **9,80** |
| **ismeretlen** | **77** | **2,74** | **240** | **2,50** |
| **Összesen** | **2806** | | **9601** | |

A táblázatból kitűnik, hogy a bizonytalan és vitatott minősítések felszámolása két új jövevényszó-kategóriának a megjelenésével is járt. Igaz, ezeknek csak egy-egy képviselőjük van, 1-1 adattal: a kaukázusi jövevény az I. statisztika szerint bizonytalan eredetű *réz,* a francia pedig az I. statisztikában a vitatott eredetűekhez tartozó *kilincs.*

Az átrendezés miatt a belső keletkezésű kategóriák sora is gyarapodott eggyel: a becéző szóalkotással. Egyetlen képviselője van, de 14 adattal: az I. típusú statisztikában a vitatottakhoz számító *bátya.*

A II. statisztikát létrehozó átminősítések az alapnyelvi örökség főkategóriájának nagyobb mértékben kedveztek, mint a belső keletkezésűeknek és a jövevényszavaknak: az örökséghez tartozók részesedése a szóállományban 10% fölé, a szóelőfordulásokat tekintve pedig 25% fölé került. Ezen nem is csodálkozhatunk, hiszen az I. statisztikával kapcsolatban láthattuk, hogy a leggyakoribb bizonytalan és vitatott eredetű szavaknak éppen a finnugor minősítésük vehető leginkább számításba. – Így persze az sem véletlen, hogy a II. típusnak az alapnyelvi rétegek közül éppen a finnugor a fő kedvezményezettje, különösen az előfordulásokra nézve: részesedése itt önmagában is 14%-ra nő.

Az ismeretlen eredetűek a II. statisztikában főkategóriává válnak, de részesedésük az I.-höz képest lényegében nem módosul.

**4. A Toldi lemmáinak eredetstatisztikája.** Ahogy arról a **2.** pontban már szó esett, a lemmastatisztikát az igenevek képzőinek, valamint a ható és a műveltető igék képzőjének a levágásával készítettem elő; szenvedő igével itt nem volt dolgom. – A művelet persze nem érinthette a melléknevesült, a főnevesült vagy más szófajváltáson átment igeneveket, hiszen ezek már az alapigétől elszakadt,

önálló szótári egységek: például *füstokádó, döglött; kopó* 'vadászkutya', *üldöző; múlva.* A műveltető képzővel létrejött, de lexikalizált *sért, veszt*-féle igékről sem metszettem le a képzőmorfémát.

Az átalakítás többféle következménnyel járt. Egyfelől eltüntette a szóállomány igeneveit, továbbá ható és műveltető igéit. Másfelől ezek helyébe vagy új szereplőként lépett az alapigéjük, vagy a korpuszban már eleve meglévő alapige adatait szaporították az igenevektől, illetve származékigéktől átkerülők. – A Toldi szókészletében nincsen például *tüzel* ige, *tüzelve* igenév viszont van. A lemmastatisztikához a *-ve* képzőt levágtam, így a lemmakészletben új elemként jelent meg a *tüzel.* – A *jár* ige önmagában is megvan a Toldi szavai között, 7 adattal. Van azonban *járhat* és *járván* is, 1-1 adattal. A ható ige és a határozói igenév a szókészletből a lemmakészletbe nem került át, viszont adatszámuk nem veszett el: a lemmák között gyarapítja a *jár* adatait, így ennek a lemmastatisztikában 7 helyett 9 adata van.

Mindebből következik, hogy a lemmák állománya természetesen kisebb lett a szavakénál, viszont a szóelőfordulásoknak és a lemma-előfordulásoknak a száma megegyezik. A Toldi lemmaállománya 2672 tagú (a szavaké 2806), 9601 előfordulással. Egy lemmára tehát 3,59 előfordulás jut (míg egy szóra 3,42).

A Toldinak az I. típusú, azaz a bizonytalanság és vitatottság kategóriáit figyelembe vevő lemmaeredet-statisztikáját a 3. táblázat foglalja össze.

**3. táblázat**

A Toldi I. típusú lemmaeredet-statisztikája

| **Eredet** | **Lemma** | **%** | **Előfordulás** | **%** |
|---|---|---|---|---|
| uráli | 43 | 1,61 | 392 | 4,08 |
| finnugor | 75 | 2,81 | 842 | 8,77 |
| ugor | 35 | 1,31 | 163 | 1,70 |
| tisztázatlan rétegből | 18 | 0,67 | 156 | 1,62 |
| kétforrású örökség | 1 | 0,04 | 21 | 0,22 |
| örökség magyar képzővel | 72 | 2,69 | 390 | 4,06 |
| **örökség összesen** | **244** | **9,13** | **1964** | **20,46** |
| származék | 605 | 22,64 | 1044 | 10,87 |
| fiktív tőből | 74 | 2,77 | 164 | 1,71 |
| képzőcserés | 1 | 0,04 | 1 | 0,01 |
| elvonás | 14 | 0,52 | 22 | 0,23 |
| szórövidülés | 9 | 0,34 | 18 | 0,19 |
| szóösszehúzás | 5 | 0,19 | 13 | 0,14 |
| jel- vagy ragszilárdulás | 220 | 8,23 | 986 | 10,27 |
| önállósulás | 4 | 0,15 | 18 | 0,19 |
| szófajváltás | 57 | 2,13 | 1255 | 13,07 |

| | | | | |
|---|---|---|---|---|
| jelentéselkülönülés | 2 | 0,07 | 3 | 0,03 |
| szóhasadás | 13 | 0,49 | 151 | 1,57 |
| ikerszó | 4 | 0,15 | 4 | 0,04 |
| összetétel | 865 | 32,37 | 1728 | 18,00 |
| szóösszevonás | 3 | 0,11 | 11 | 0,11 |
| tükörösszetétel | 1 | 0,04 | 1 | 0,01 |
| onomatopoetikus | 63 | 2,36 | 150 | 1,56 |
| belső, de tisztázatlan | 20 | 0,75 | 204 | 2,12 |
| **belső keletkezésű összesen** | **1960** | **73,35** | **5773** | **60,13** |
| permi | 1 | 0,04 | 6 | 0,06 |
| iráni | 9 | 0,34 | 37 | 0,39 |
| török | 62 | 2,32 | 204 | 2,12 |
| német | 32 | 1,20 | 73 | 0,76 |
| latin | 15 | 0,56 | 28 | 0,29 |
| olasz | 13 | 0,49 | 22 | 0,23 |
| román | 1 | 0,04 | 2 | 0,02 |
| szláv | 98 | 3,67 | 374 | 3,90 |
| tisztázatlan nyelvből | 4 | 0,15 | 8 | 0,08 |
| közvetített | 5 | 0,19 | 8 | 0,08 |
| jövevény magyar képzővel | 7 | 0,26 | 13 | 0,14 |
| vándor | 9 | 0,34 | 23 | 0,24 |
| **jövevény összesen** | **256** | **9,58** | **798** | **8,31** |
| bizonytalan | 96 | 3,59 | 547 | 5,70 |
| vitatott | 34 | 1,27 | 269 | 2,80 |
| ismeretlen | 82 | 3,07 | 250 | 2,60 |
| **tisztázatlan összesen** | **212** | **7,93** | **1066** | **11,10** |
| **Összesen** | **2672** | | **9601** | |

A szóállományhoz képest a lemmaállományban a főkategóriákat tekintve nő az alapnyelvből örökölteknek, a jövevényszavaknak és a tisztázatlan eredetűeknek a részesedése, de mindegyiké csak 1 százalékpontnál kevesebbel. A belső keletkezésűek viszont veszítenek, de 2 százalékpontnyinál kevesebbet.

Ha a lemmastatisztika létrehozásának módjára gondolunk, egészen természetes, hogy a lemmatizáció a származékokat érinti leghátrányosabban: körülbelül 7 százalékpontot veszítenek a részesedésükből. Emiatt nemcsak búcsúzni kénytelenek az első helyüktől, hanem 10 százalékpontos hátránnyal szorulnak a második helyre az összetett szavak mögé. Az összetett szavak viszont 30%-nál jobb eredménnyel kerülnek az élre. A két kategória helycseréjében fontos szerepe van annak, hogy a 2. pontban említett *elfeledhet, előrehajolva* típusú származékszavak a lemmatizáció révén átkerülnek az összetettekhez *(el + feled, előre + hajol).*

A lemma-előfordulás okat tekintve a szóstatisztikához képest lényegében ugyanúgy változik a főkategóriák részesedése, mint az állományban: az örökségé, a jövevényeké és a tisztázatlanoké nő, a belső keletkezésűeké pedig csökken.

A lemma-előfordulások között az összetettek részesedése 1 százalékponttal nagyobb, mint a szóstatisztikában. A származékok viszont csökkenő számuk miatt a változatlan eredményű szófajváltás mögé szorulnak a kategóriák listáján.

Mivel lemmastatisztikát korábban a 18. századi sakk-könyvből és 21. századi beszélt nyelvi szövegekből készítettem (vö. HORVÁTH 2010: 427, 2016: 314–315), a Toldiét ezekkel tudom összehasonlítani. Mindkettőhöz viszonyítva ugyanolyan irányú az eltérés: a másik két korpuszban 80% feletti, azaz a Toldi-belinél nagyobb a belső keletkezésű lemmák részesedése, míg a másik három főkategória képviselete a Toldiban jelentékenyebb. Ezt most csak tényként állapítom meg, a különbség okainak feltárása külön, részletes elemzést kívánna és érdemelne.

A lemmák II. típusú, vagyis a bizonytalanság és vitatottság kategóriáit felszámoló statisztikáját az I.-höz képest ugyanúgy alakítottam ki, mint a szavak esetében. Az eredményt a 4. táblázat mutatja be.

**4. táblázat**

A Toldi II. típusú lemmaeredet-statisztikája

| **Eredet** | **Lemma** | **%** | **Előfordulás** | **%** |
|---|---|---|---|---|
| uráli | 50 | 1,87 | 413 | 4,30 |
| finnugor | 106 | 3,97 | 1385 | 14,43 |
| ugor | 46 | 1,72 | 211 | 2,20 |
| kétforrású örökség | 1 | 0,04 | 21 | 0,22 |
| örökség magyar képzővel | 100 | 3,74 | 481 | 5,01 |
| **örökség összesen** | **303** | **11,34** | **2511** | **26,15** |
| származék | 617 | 23,09 | 1112 | 11,58 |
| fiktív tőből | 94 | 3,52 | 206 | 2,15 |
| becéző | 1 | 0,04 | 14 | 0,52 |
| képzőcserés | 2 | 0,07 | 2 | 0,02 |
| elvonás | 16 | 0,60 | 24 | 0,25 |
| szórövidülés | 11 | 0,41 | 21 | 0,22 |
| szóösszehúzás | 5 | 0,19 | 13 | 0,14 |
| jel- vagy ragszilárdulás | 225 | 8,42 | 1146 | 11,94 |
| önállósulás | 4 | 0,15 | 18 | 0,19 |
| szófajváltás | 64 | 2,40 | 1263 | 13,15 |
| jelentéselkülönülés | 5 | 0,19 | 8 | 0,08 |
| szóhasadás | 15 | 0,56 | 157 | 1,64 |
| ikerszó | 4 | 0,15 | 4 | 0,04 |
| összetétel | 866 | 32,41 | 1735 | 18,07 |
| szóösszevonás | 3 | 0,11 | 11 | 0,11 |

| | | | | |
|---|---|---|---|---|
| tükörösszetétel | 1 | 0,04 | 1 | 0,01 |
| onomatopoetikus | 65 | 2,43 | 152 | 1,58 |
| **belső keletkezésű összesen** | **1998** | **74,78** | **5887** | **61,32** |
| permi | 2 | 0,07 | 7 | 0,07 |
| iráni | 15 | 0,56 | 64 | 0,67 |
| kaukázusi | 1 | 0,04 | 1 | 0,01 |
| török | 78 | 2,92 | 286 | 2,98 |
| német | 37 | 1,38 | 88 | 0,92 |
| latin | 18 | 0,67 | 34 | 0,35 |
| francia | 1 | 0,04 | 1 | 0,01 |
| olasz | 15 | 0,56 | 31 | 0,32 |
| román | 1 | 0,04 | 2 | 0,02 |
| szláv | 102 | 3,82 | 391 | 4,07 |
| jövevény magyar képzővel | 10 | 0,37 | 25 | 0,26 |
| vándor | 10 | 0,37 | 24 | 0,25 |
| **jövevény összesen** | **290** | **10,85** | **954** | **9,94** |
| **ismeretlen** | **81** | **3,03** | **249** | **2,59** |
| **Összesen** | **2672** | | **9601** | |

A lemmatizáció (a II. szóeredet-statisztikához viszonyítva) itt is csak a belső keletkezésűek főkategóriájának okozott veszteséget, és csupán 1 százalékpontnyit.

A legnagyobb vesztes a kategóriák közül itt is a származékoké. Az előfordulásokat tekintve a belső keletkezésűek közül az összetételen kívül a szófajváltás és a szilárdulás is gyakoribb nála.

Az előfordulásokra nézve kiemelést érdemelnek a finnugor elemek: akárcsak a II. szóeredet-statisztikában, a lemmák között is csak az összetettek állnak előttük.

**5. Összegzés.** Nem állíthatom, hogy a Toldi szókészletének etimológiai statisztikai vizsgálata nagy meglepetésekkel járt. Úgy vélem, mégsem volt hiábavaló az elvégzése, mivel eredményei – mind a korpusz terjedelmének köszönhetően, mind Arany Jánosnak és Toldijának a jelentőségére tekintettel – tájékozódási pontul, összehasonlítási alapul szolgálhatnak más hasonló tárgyú felmérésekhez, és általában véve is fogódzót nyújthatnak az etimológiai kérdésekben való eligazodáshoz.

Ne feledjük: a tőstatisztika elkészítése még hátra van, a kép azzal válik majd teljessé.

Biztos vagyok abban, hogy a Toldi etimológiai statisztikájának számszerű eredményei és módszertani tanulságai az oktatásban is hasznosíthatók. Hogy annak a különböző szintjein mely részleteket és milyen szempontból érdemes kiemelni, annak eldöntésében természetesen elsősorban a tanárok, oktatók az illetékesek.

**Kulcsszók:** állomány és előfordulás, Arany-év, etimológiai statisztika, szóstatisztika és lemmastatisztika, Toldi.

## Hivatkozott irodalom

Beke József 2017. *Arany-szótár* 1–3. Anyanyelvápolók Szövetsége – Inter, Budapest.
Benkő Loránd 1962. Adatok a magyar szókincs szerkezetének változásához. *Nyelvtudományi Közlemények* 64: 116–136.
Csiszár Gábor 2002. *Kosztolányi első verseskötetének etimológiai vizsgálata.* Szakdolgozat. Kézirat. ELTE, Budapest.
EWUng. = *Etymologisches Wörterbuch des Ungarischen* 1–2. Hrsg. Benkő, Loránd. Akadémiai Kiadó, Budapest, 1993–1995. + *Register.* Akadémiai Kiadó, Budapest, 1997.
Horváth László 2000. Etimológiai kategóriák arányai mai elbeszélésekben. *Magyar Nyelv* 96: 170–181, 316–332.
Horváth László 2002. Az Ómagyar Mária-siralom etimológiai statisztikája. *Magyar Nyelv* 98: 265–282.
Horváth László 2004. *Két Halotti beszéd az etimológia tükrében.* Előadás. Kézirat. MTA Nyelvtudományi Intézet, 2004. 02. 26.
Horváth László 2010. Az első magyar sakk-könyv etimológiai statisztikája. *Magyar Nyelvőr* 134: 421–436.
Horváth László 2012. *Három Halotti beszéd az etimológiai statisztika tükrében.* Előadás. Kézirat. MTA Nyelvtudományi Intézet, 2012. 11. 20.
Horváth László 2016. Nemzedékek spontán beszéde etimológiai megközelítésben. In: Balázs Géza – Veszelszki Ágnes szerk., *Generációk nyelve.* ELTE BTK Mai Magyar Nyelvi Tanszék – Inter – Magyar Szemiotikai Társaság, Budapest. 309–317.
ToldiSz. = Pásztor Emil 1986. *Toldi-szótár. Arany János Toldijának szókészlete.* Tankönyvkiadó, Budapest.
Tolnai Vilmos 1924. Halhatatlan magyar nyelv. *Magyar Nyelv* 20: 50–59.
Zsilinszky Éva 2003. Az újabb magyar kor. Szókészlettörténet. In: Kiss Jenő – Pusztai Ferenc szerk., *Magyar nyelvtörténet.* Osiris Kiadó, Budapest. 804–823.

## On the origin of the word stock of Arany's Toldi

The author of this paper, one of the contributors to Etymologisches Wörterbuch des Ungarischen, has been doing etymological statistics for almost 20 years now. Several of his publications discuss present-day Hungarian texts and/or documents of various genres from earlier periods in that respect. This time, he commemorates the 200 years anniversary of János Arany, the writer of the most widely know Hungarian epic poem, Toldi, by presenting an etymological statistics of that poem. The author explores the ratios of various categories of origin in the word stock of the poem, both in terms of types and tokens. The results are compared to those of the author's earlier investigations. All that is done in the hope that this analysis will be able to serve both as a baseline for future studies and as educational material from primary schools to PhD courses.

**Keywords:** Arany anniversary, etymological statistics, word statistics and lemma statistics, Toldi, type and token.

Horváth László
MTA Nyelvtudományi Intézet