JUHÁSZ ZOLTÁN
kutatómérnök, népzenész, Központi Fizikai Kutatóintézet

A GYÖKRENDSZER SZÁMÍTÓGÉPES VIZSGÁLATA

1. BEVEZETÉS

A Czuczor–Fogarasi-szótár alapeszméjének – a magyar nyelv gyökrendszerének – megismerése sokunk számára anyanyelvünk, sőt nem túlzás azt mondani: a világ újrafelfedezésének élményét adta. A villamos ablakán kibámulva, séta közben, vagy valamely hivatalban a sorunkra várva ízlelgetjük a szavainkban rejlő kapcsolatokat, csodálkozunk rá ezeken keresztül a világ olyan összefüggéseire, melyeket már ezerszer kimondtunk, mégsem ismertünk fel. Így a gyökrendszer a nyelvészettől távoli tudományok számos művelőjének gondolkodását is megtermékenyítette áttekinthető, szemléletes világképével: a nyelv univerzumában elkülönülő, vagy összekapcsolódó szó-galaxisok mintegy leképezik a világ összetartozó és össze nem tartozó dolgait.

E sorok írója már régen gondolkozik azon, miképpen lehetne láthatóvá tenni ezt a nyelv-univerzumot. Ez nem csupán egy látványos számítógépes játék kieszelését jelenti, mivel a kérdést úgy is feltehetjük: megadható-e olyan **algoritmus**, mely valóban szóbokrokba – szó-galaxisokba – képezi le szavainkat? Ha találunk ilyet, az súlyos érv a gyök-elmélet mellett: bizony leképezhető a nyelv a gyökök köré szerveződő szóbokrok áttekinthető rendszerébe. Az alapötletet végül a dallamok bokrainak vizsgálata közben szerzett tapasztatok adták:

népdalok rokonsági rendszerének elemzésére ui. kidolgoztam már az adatbányászat számos területén alkalmazott, ún. MDS algoritmus egy változatát, mely a dallamokat képviselő pontokat a zenei „távolságoknak" megfelelően rendezi el a térben vagy a síkon (Borg 2005; Juhász 2011). A feladat tehát **a szavak „távolságait" kiszámító algoritmus** kidolgozása. Ha ez megvan, az MDS algoritmus megmutatja, valóban bokrokba szerveződnek-e szavaink, vagy valamilyen más rendszerbe.

Szavak távolságának jellemzésére már számos eljárást kidolgozott a **számítógépes nyelvészet** (Cohen 2003; Kohonen 1998). Ezek között az egyik legelterjedtebb az ún. Levensthein-féle eljárás, így első lépésként én is ezt alkalmaztam a magyar szavak távolságának kiszámítására. Az eredmény – mint látni fogjuk – a várakozásokon alulinak bizonyult, mind a távolságértékek valósághűsége, mind a kapott összkép tekintetében (Navarro 2001). Szükség mutatkozott tehát egy új, általánosnak és ésszerűnek tekinthető saját távolságmérő eljárásra: olyanra, ami a szavak leghosszabb közös szakaszát keresi, így „észreveszi" ugyan a közös gyököket is, de az egyszerű gyökkeresésnél jóval általánosabb.

A bevett alkalmazások egy része, köztük a Levensthein-módszer is csak a két összehasonlítandó szóra figyel. Létezik azonban egy másik irányzat is, az ún. **korpusznyelvészet**: ez nagy szövegtestek elemzése alapján igyekszik megtalálni az összefüggő szövegeket, vagy gyakran együtt előforduló, így valószínűleg azonos fogalomkörbe tartozó szavakat (Mc Eney 1996; Terra 2003). Mások egész nyelvrokonsági fákat számítottak ki szövegtestek elemzésével (Benedetto et al. 2002; Bencze 2002). A magyar nyelv számítógépes morfématárának kidolgozása a legújabb eredmények közé tartozik, miközben a gyökkorpusz létrehozását is megkezdték Kresznerics Ferenc módszerét követve (Kiss 2011).

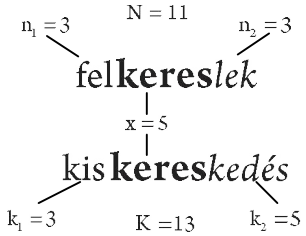
Az itt bemutatandó módszer is egy **szövegtest statisztikai elemzése** alapján értékeli a szavak rokonságát, így e tekintetben a korpusznyelvészeti irányzathoz tartozik. A szövegtest az *Újtestamentum* egy

része: *János Evangéliuma, Az Apostolok cselekedetei, Pál levelei a rómaiakhoz és a korinthusiakhoz.* Ezt a szövegtestet számos nyelv bibliafordításaiból összegyűjtöttem, hogy a nyelvek közti összehasonlítást közös alapra helyezzem. A szövegek mérete 300 000–400 000 leütés között mozog. A számítógépes nyelvészet egyik legfontosabb területe a párhuzamos szövegek statisztikai elemzésére épülő fordítóprogramok fejlesztése, így jó alappal bízhatunk abban, hogy a módszer a különböző nyelvekben meglevő azonos jelentésű gyökök szóbokrainak összevetésére is alkalmas (Laki 2010).

A módszer eredményességét a szövegben előforduló szavak bokrainak az algoritmus által megszerkesztett ábráival igyekszem bizonyítani. Mivel pedig a CzF. következetesen felsorolja a magyar szavak másnyelvi rokonait, összehasonlításul ezek bokrait is bemutatom, sőt, *A magyar nyelv történeti-etimológiai szótára* adataival (TESz.) is összevetem. Nem lévén nyelvész, kénytelen vagyok vállalni a tévedés kockázatát, de úgy gondolom, ha sikerül felkelteni a nyelvészek érdeklődését, a kockázat már megérte.

2. A TÁVOLSÁG DEFINÍCIÓJA

Távolságdefiníciónk bevezetéséhez az 1. ábrán látható példát használjuk szemléltetésül.



1. ábra

A szavak közös szakaszának keresése és eltérésük mérőszámai

Először is megkeressük az összehasonlítandó két szó leghosszabb közös szakaszát: ez a „felkereslek" és a „kereskedés" szavak esetén: a „keres". A szavak eltérését elsődlegesen a kimaradó részek hossza jellemzi: esetünkben a „fel", „lek", ill. „kis" és „kedés" hosszai. Minél nagyobbak ezek a teljes szóhosszakhoz (N és K) képest, annál kevésbé hasonlít egymásra a két szó. Ezt a követelményt legegyszerűbben az

$$\frac{n_1 + n_2}{N} \cdot 0{,}5 + \frac{k_1 + k_2}{K} \cdot 0{,}5$$

képlet fejezi ki. Az eredmény 1, ha a maradékok összhossza a szavak hosszával azonos (vagyis ha egyáltalán nincs közös szakasz), és 0, ha egyáltalán nincs maradék (vagyis ha a két szó azonos).

Nem mindegy azonban, hogy a közös rész, ill. a maradékok értelmes szavak-e. Nyilvánvalóan közelebbi kapcsolatot szeretnénk mérni akkor, ha a közös rész magában is értelmes: a „keres" esetében ez teljesül is. Ellenpéldánk legyen a „rókának" és az „apókát" szópáros: itt a leghosszabb közös rész az „óká", ez pedig önmagában értelmetlen, összefüggésben azzal, hogy a két szónak valóban nincs fogalmi kapcsolata egymással. Szeretnénk, hogy ez a tény a szavak távolságában számszerűleg is kifejeződjön. Ehhez a fenti képletet meg kell még szorozni egy 0 és 1 közötti súlytényezővel, mely legkisebb értékét akkor veszi fel, ha a közös szakasz biztosan értelmes, a legnagyobbat pedig akkor, ha biztosan értelmetlen. Az algoritmus az értelmesség kérdését csak úgy tudja eldönteni, ha rendelkezésére áll egy szókészlet, és ebben megkeresheti a közös szakaszt: ha nem találja, akkor a súlytényező legnagyobb értékével számol, ha viszont megtalálja, akkor kisebbel. Távolságdefiníciónk továbbfejlesztése emiatt igényli minél nagyobb szókészlet használatát. Elvileg ez egy teljes szótár is lehetne. A szótárban azonban nincsenek meg, vagy csak hiányosan vannak meg a szavak toldalékolt származékai, így még egy teljes szótár sem biztosítaná, hogy algoritmusunk az összes értelmes hangsort azonosítani tudja. További probléma, hogy ha vizsgálatainkat több nyelvre is ki kívánnánk terjeszteni, a kérdéses nyelvek összes digitalizált szótárára szük-

ségünk lenne. Ezekben a nagy (ám szempontunkból mégsem teljes) adatbázisokban nagyon időigényes lenne a keresés, hiszen az összes létező alapszón – a legritkábban előfordulókon is – minden esetben végig kéne menni. Ezért inkább a korpusznyelvészet útját követjük: egy kellően nagy **szövegtestből** kigyűjtünk minden szót (a toldalékoltakat, összetetteket, igekötőseket, stb. is, válogatás nélkül), és ezeket **hosszuk szerint** rendezve tároljuk, **előfordulási gyakoriságaikkal** együtt. Minél reprezentatívabb a szövegtest, annál pontosabban működik algoritmusunk.

Mielőtt definíciónkat kiegészítenénk a szövegtestből származtatott „értelmességi" súllyal, gondoljuk meg, hogy az értelmesség vizsgálata a maradék részekre is kiterjeszthető, és ezzel további hibás rokonításokat kerülhetünk el. Például a „tímár" és a „szamár" közös része az értelmes „már", az összerendelés mégis esetleges. Erre az algoritmus maga is „rájöhet", ha a maradékokat („tí" és „sza") is keresi a szókészletében. Nyilván nem találja őket, és ezt a tényt újabb súlytényezők magas értékével ki is fejezi.

Mindezen meggondolások alapján a szavak távolságát a következő képlettel definiáljuk:

$$t = \left(\frac{s_{1,1} \cdot n_1 + s_{1,2} \cdot n_2}{N} \cdot w + \frac{s_{2,1} \cdot k_1 + s_{2,2} \cdot k_2}{K} \cdot (1 - w)\right) \cdot S, \qquad (1)$$

Ahol S a leghosszabb közös szakasz, pedig a maradékok „értelmességét" jellemző súlytényezők. A w tényező a rövidebb, ill. hosszabb szó részesedését határozza meg az eredő távolságban (súlyozott átlagolás).

A **súlytényezők** kiszámítása a szókészlet és benne a szavak előfordulási gyakorisága alapján történik, és elvileg annak bármilyen monoton csökkenő függvénye lehet. Mi az $1/(előfordulás + 1)$ képlettel számolunk, vagyis példánkban:

$$s_{1,1} = \frac{1}{előford\ („fel") + 1} \qquad s_{1,2} = \frac{1}{előford\ („lek") + 1}$$

$$s_{2,1} = \frac{1}{előford\ („kis") + 1} \qquad s_{2,2} = \frac{1}{előford\ („kedés") + 1}$$

$$S = \frac{1}{előford\ („keres") + 1} \qquad (2)$$

Azt, hogy a számítás milyen mértékben képes megkülönböztetni egymástól az értelmes és az esetleges alaki egyezéseket, a két példa is jól mutatja. Az értelmes közös és maradék szakaszok előfordulásaira egyaránt 9-et, az értelmetlenekre 0-t feltételezve a „felkereslek – kereskedés", ill. a „rókának – apókát" távolságra ~0,036, ill. 0,68 adódik, vagyis az esetleges megfelelés több mint egy nagyságrenddel nagyobb távolságot is adhat, mint az értelmes. Ennek alapján kimondhatjuk, hogy **távolság-definíciónk** a szavakat nem egyszerűen csak **alaki hasonlóságuk**, hanem **jelentésük** kapcsolata alapján is rokonítja.

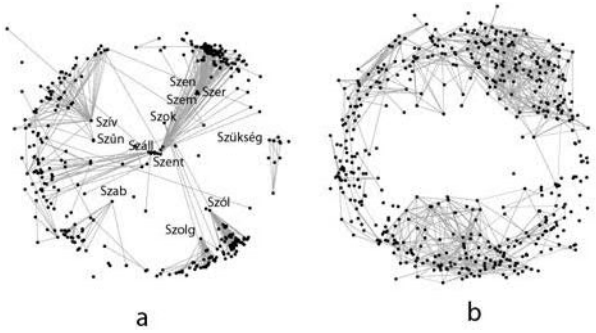
3. A SZAVAK KAPCSOLATAINAK ÁBRÁZOLÁSA ALACSONY DIMENZIÓJÚ TEREKBEN

Az 1–2. képletek alapján a szövegtestünkben előforduló összes szó egymás közti távolságait kiszámíthatjuk. Felmerül a kérdés, hogyan tekinthetnénk át **e hatalmas szókészlet** (a magyar szövegben pl. ~11 000 féle szó van. A továbbiakban következetesen két szóköz közötti hangsort tekintünk szónak.) **rokonsági viszonyait**. A kérdés már csak azért is izgalmas, mert távolságdefiníciónkban igyekeztünk a szavak értelmi rokonságáig is eljutni – így az eredménytől valamilyen „értelmezési térben" való rendeződést várunk.

A megoldást az adatbányászat irodalmában angol rövidítéssel MDS-ként (multidimensional scaling) szereplő, magyarul talán „önszer-

vező hangyabolynak" nevezhető algoritmus kínálja. Ennek bemenete lehet a szavaink távolságait tartalmazó (M db. szó esetén M*M-es méretű) szimmetrikus mátrix. **Az algoritmus** M db. pontot igyekszik elhelyezni egy síkon, vagy egy háromdimenziós térben oly módon, hogy a pontok távolságai a lehető legjobban megfeleljenek a bemenő mátrix adatainak – esetünkben tehát a szavak 0 és 1 közötti távolságértékeinek (Borg, 2005, Juhász, 2011).

A szövegtest „sz" betűvel kezdődő szavait pl. az algoritmus a 2a ábrán látható módon rendezte el. Ezen jól látható, hogy a szavak túlnyomó többsége valamilyen **háromhangos gyökszó** környezetébe tömörül, abból valóban úgy sarjad ki, mint a bokros növény a gyökeréből. (Az önállóan nem létező gyököket, mint pl. a „szok", vagy a „szűn" – a rendszer egy tanulási folyamat során „bányászta" ki a szövegből. Az öntanuló algoritmust azonban itt nem tárgyaljuk.) Fontos hangsúlyozni, hogy ezt az elrendeződést semmilyen „külső előprogramozás" nem segítette: távolságdefiníciónk mindig a lehető leghosszabb közös szakaszokat keresi, és nem tünteti ki a háromhangosakat. Az ábra tehát bízvást tekinthető a nyelvi valóság „fényképének" – igaz, csak olyan fényképnek, melynek „optikáját" az 1–2. képletek távolságdefiníciója adta. Hogy az „optika" mennyit számít, azt a 2b ábrával való összevetés mutatja: a számítógépes szövegelemzésben elterjedten használt „normált Levensthein"-féle távolságdefiníció egyáltalán nem ilyen „gyökös" elrendeződésbe képezi le ugyanazokat a szavakat (Navarro, 2001).

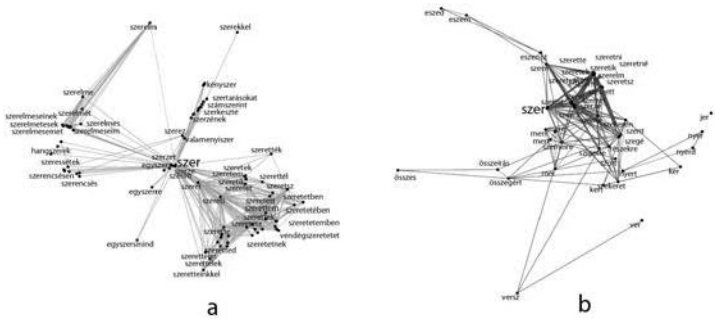


2. ábra
a: *sz*-szel kezdődő szavak rendeződése saját távolságértékeink alapján; b: ugyanezen szavak elrendeződése normált Levensthein-távolság esetén

A tapasztalat szerint azonban a mi távolságdefiníciónk jobb találati arányt mutat. Ennek érzékeltetésére vessük össze a 3a és 3b ábrákat! Ezeken már csak a teljes szókészletet ábrázoló 2. ábra egy kis részletét, a „szer" gyök környezetét látjuk. Az első ábra a gyök bokrát a mi távolságdefiníciónk alapján mutatja, a második a normált Levensthein-féle távolsággal kigyűjtött bokrot. Most tehát a két szókészlet nem szükségszerűen azonos. Valóban, a Levensthein távolság rátalált ugyan a „szer-szeret-szeretné", stb. típusú bokor rövidebb szavaira, de már a hosszabbakat (pl. „vendégszeretetet" – lásd 3a ábra) nem találta meg. Ugyanakkor számos hibás találatot mutat a rövid szavak között (*összes, eszem, versz*, stb. ld. 3b ábra.). A hibás találatokat a továbbiakban sem javítjuk ki, mivel a tanulmány egy **számítógépes eljárás működését** kívánja bemutatni, ehhez pedig a statisztikai módszerek esetében szükségszerűen fellépő **hibák** is hozzátartoznak. (Saját módszerünk anyanyelvi értékelése céljából az 1–2. képletek szerinti, teljes szövegtestre kiterjedő gépi kereséssel kilistáztuk az összes háromtagú szó bokrait: a találatok hét személy független értékelése alapján megközelítőleg 80%-ban bizonyultak helyesnek.)

A 3a ábra jóval rendezettebb képén szépen elkülönülnek a „szeret" és a „szerelm-es", stb. rész-felhők, és ezek a hosszú szavakat is magukba foglalják (pl. „szerelmeseimet").

A gyök számos más sarjadéka is ott van az ábrán, mivel azonban ezek a szövegtestben nem alkotnak akkora bokrokat, mint az előbbi kettő, zömükben egy vonalszerű nyúlványba rendeződnek (*szerez, szerkeszt, szám szerint, egyszersmind* stb.), így adva helyet az *Újtestamentum* kulcsfogalmának, a „szeret", „szerel(e)m" nagy bokrainak. A leképezés egyébként jóval differenciáltabb képet ad három dimenzióban, ez azonban csak a képernyőn való forgatással szemléltethető.



3. ábra
a: a „szer"- gyök bokra saját távolságértékek alapján
b: a „szer"- gyök bokra Levensthein távolságok alapján

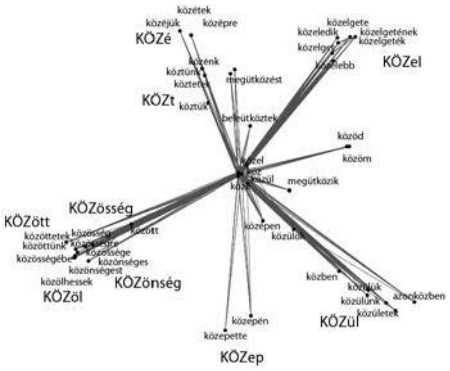
A gyök hatalmas bokrára tekintve egész természetesnek tűnik pl. a „szerencsés, szerencsésen" szavak jelenléte is. Valóban, a CzF. szótár is a „szer" gyökből származtatja a „szerencse" szót, amellett, hogy annak szláv, mongol és szanszkrit kapcsolatait is elemzi. A TESz. ezzel szemben már egyértelműen a szláv eredet mellett van: a szerb-horvát *sreća*, szlovén *sreča* ('sors, szerencse') szavakból eredezteti. Szerinte a szláv szavak eredeti jelentése '(össze)találkozás', ebből lett 'az események találkozása', azaz 'véletlen'. Érdemes tehát megvizsgálnunk, vajon mekkora bokra van a horvát szónak saját nyelvében.

Nos, a *„sreća"* a magyarnak pontosan megfelelő horvát Biblia-szakaszban egyetlen rokonnal (*„sreće"*) rendelkezik, és így van ez a *„sreć"*- csel, sőt még a *„sre"*-vel is. Természetesen nem minden nyelv alkalmas egyformán szóbokrok fejlesztésére, ám a hatalmas magyar bokor olyan logikusan és értelmezhetően foglalja magába a *szerencse* szót is, hogy a szlovén szó bokortalanságától függetlenül is kijelenthetjük: Czuczorék érvelése és a gépi elemzés ez esetben kölcsönösen erősítik egymást, és a „szerencse" „szer"-ből való származtatását támogatják a szláv eredettel szemben.[1]

4. NÉHÁNY SZÓBOKOR ÉS KAPCSOLATAIK

A továbbiakban éppen a fenti módszert igyekszünk követni. Egyenként bemutatjuk néhány, a szövegtestben előforduló, háromtagú szavunk bokrát, és ahol a rendelkezésünkre álló bibliafordítások lehetővé teszik, a velük rokonított idegen szavak bokrait is bemutatjuk. A rokonításokat kizárólag a CzF.-ből és a TESz.-ből vesszük, tehát technikánkat csakis nyelvészek által már kimondott kapcsolatok elemzésére használjuk, nem kísérletezünk más, „délibábos" szóegyezések felkutatásával. Következtetéseinket főképpen a szóbokrok méreteire és szerkezetére igyekszünk alapozni, mintegy „gépesítve" a Czakó Gábor által már sok esetben – köztük a következő példák jó részében is - alkalmazott módszert (Czakó 2008, 2010).

Módszerünk képes értelmileg is elkülöníteni a rokon szócsoportokat, amennyiben az értelmi sajátságoknak szabályos formai megnyilvánulásaik is vannak. Ezt mutatja be a 4. ábra.



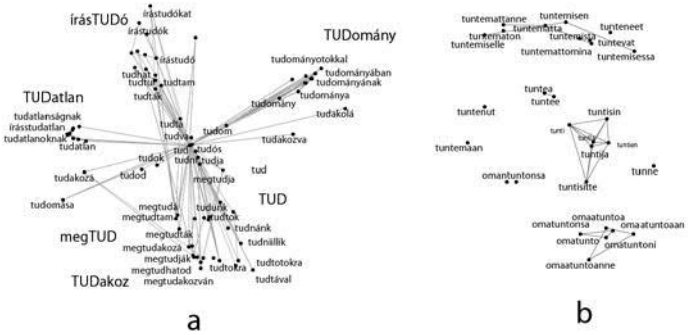
4. ábra
A „köz" szó bokra

A „köz" sarjadékai értelmük szerint is igen jól elkülönülő nyalábokban nőnek ki a gyökből. Elválik a „közep", „közül", „közel", „közé" család. A „között", „közöl", „közösség", „közönség" családok viszont a közös második *-ö* hang miatt egy nyalábba vegyülnek. Ez mutatja, hogy az algoritmus „értelmezési" képessége valójában **formai jegyekre** támaszkodik, és csak akkor lehet sikeres – akkor viszont lehet –, ha az értelmi különbségek formai különbségekben is megjelennek. A három- vagy négydimenziós leképezés a tapasztalat szerint sokat javíthat a felbontáson, de általában csak a képernyőn forgatva válik látványossá.

Módszerünk természetesen a **névszók és igék ragozott alakjait** is feltünteti a bokorban, így az értelmileg nagyon hasonló szavak is növelik annak méretét – pl. „közönséges", „közönségest". Ezt a jelenséget automatikusan kiszűrni csak úgy lehetne, ha bizonyos nyelvtani szabályokat előre beprogramoznánk, pl. rangsorolnánk a toldalékokat fogalommódosító erejük szerint. Mi azonban egyelőre inkább ragaszkodnánk a távolságdefiníciónkban megfogalmazható matematikai elvek tiszta érvényesüléséhez, és a felülbírálást a nyelvész szakemberre bíznánk. Mindenesetre tény, hogy a bokor mérete mellett annak fogalmi gazdagsága is figyelembe veendő, ám ez utóbbi gépi mérésére

egyelőre nem törekszünk. A „köz”-t a TESz. manysi, hanti és mari szavakkal rokonítja, e nyelvek bibliafordításait azonban eddig nem állt módunkban feldolgozni.

A „köz”-nél is gazdagabb a „tud” bokra a vizsgált szövegtestben (5. ábra). Az értelmi szétválás azonban még ilyen „zsúfoltság” mellett is jól kivehető: szépen elkülönülnek a „tudomány”, „tud”, „tudakoz”, „megtud”, „tudatlan”, „írástudó” nyalábjai, és magányosan még a „tudós”, „tudtul” is megjelenik. A CzF. török, finnugor, perzsa, valamint a TESz. finn rokonításai közül a finn „tuntea” (tudni) létszámában és fogalmi gazdagságában is hasonló bokrát sikerült kimutatni (*tuntee* – ’érzés’; *omatunto* – ’lelkiismeret’;[2] *tuntematon*[3] – ’ismeretlen’ stb.). Az ábrán kisebb betűkkel írott hibás találatokat (*tunti*, *tuntia*, *tuntien* – Pomozi Péter közlése) természetesen nem töröltük utólag az algoritmus által szerkesztett gráfból.



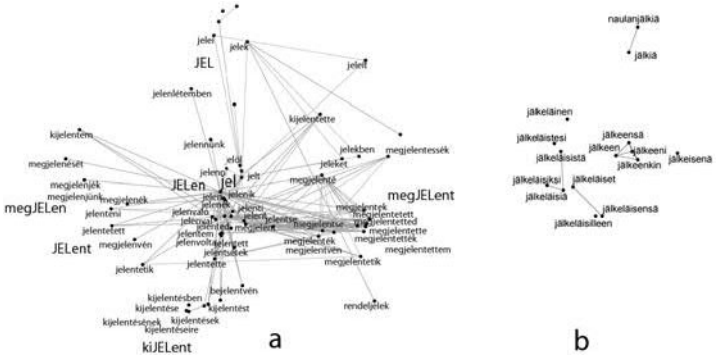
5. ábra
a: a „tud” szó bokra a magyar szövegben
b: a „tuntea” szó bokra a finn szövegben

A „tud” és a „tuntea” magyar és finn bokrai tehát a TESz. rokonításaival összeegyeztethető formai és fogalmi megfeleléseket mutatnak, így azokkal kölcsönösen erősítik egymást.

A magyar „jel” jelentős szóbokorral van jelen a vizsgált szövegben (6. ábra). Az első pillantásra meglehetősen kusza ábrát jobban szem-

ügyre véve, mégis kivehetők az elkülönülő „jel", „jelen", „jelent", „megjelent", „kijelent", „megjelenik" alapjelentésű családok. (*jel, jelöl, jelenik, jelent, jelenvaló, megjelenés, kijelent, bejelent* stb.) A „jel" a TESz. szerint, de a CzF. szerint is finnugor kapcsolatú szavunk, mely a finnben *jälki* 'lábnyom' formában él. A finn szó bokra kisebb a magyarnál, és ismét szétszóródott képet mutat, a bokor tagjait pedig kivétel nélkül a térbeli, vagy időbeli követés értelme kapcsolja össze (*jälki* – 'következő, lábnyom'; *jälkeen* – 'után'; *jälkeläinen* – 'utód'; *naulenjälkiä* – 'szegek helye').[4]

Ez a jelentésbokor véleményem szerint tartalmilag csak igen szűk mezsgyén kapcsolódik a magyar jelhez (a lábnyom valóban jele annak, hogy valaki ott járt). Ugyanakkor azt sem mondhatjuk, hogy a finn bokor fogalomköre a magyar egy szűk része volna csupán. Nem szűkebb, hanem más, mely csak egy lehatárolt értelmezési tartományban van átfedésben a magyarral.



6. ábra
A magyar „jel" (a) és a finn „jälki" ('nyom' b) bokrai
a megfelelő szövegekben

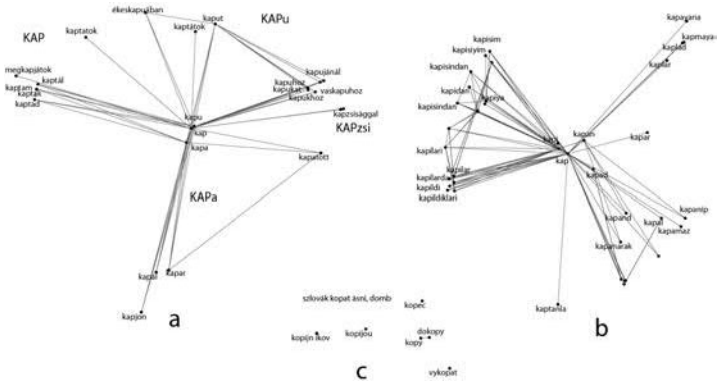
Tekintsünk most egy olyan példát, ahol a rokon szóbokor méretében is, jelentéseiben is jól megfelel a magyarnak. A „kap" gyökünk bokrában a „kap" „kapu", „kapa", „kapzsi" szavakat és toldalékolt sarjaikat

fedezhetjük fel (7. ábra). A CzF. szerint ezek a szavak nem mind tartoznak össze, mivel a „kapu" a kaparást kifejező alvó gyök származéka, a „megkap, elkap" viszont egy másik „kap" gyöké. Kétségtelen, hogy módszerünk **nem képes** megkülönböztetni alakilag teljesen megegyező, ám más jelentésű szavakat. Kérdés azonban, hogy valóban ilyen helyzettel állunk-e szemben, vagyis hogy *összekapcsolható*-e a „kaptad" ige a „kapu" főnévvel. A CzF. a *kap* élő igei formájához török, finn, továbbá latin, német, szanszkrit és görög párhuzamokat sorol fel. A „kap" (tároló, edény, kupa?)[5] török gyök bokrát megfejtve, egyaránt találkozunk 'megragad (elkap), csapda (elkapó)', ill. 'kapu, fedett, zárt, zárás, befed' jelentésű szavakkal: (*kap* – 'edény'; *kapi* – 'ajtó, kapu'; *kapmaya* – 'kiragad, megragad'[6]; *kapan* – 'csapda, háló'[7]).

Mint már jeleztük, az ábrák a gépi elemzés esetleges hibáit is tartalmazzák, de ez nem akadályoz minket a fontosabb jelentéstartományok vizsgálatában. A méretes török bokorban eszerint ugyanúgy együtt vannak az 'elkapás'és a 'lezáró, befedő kapu' jelentés különböző fejleményei, mint magyar megfelelőjében. Ez pedig arra utal, hogy a magyar bokor ugyanúgy egyetlen gyök sarjadéka, mint a török. A CzF. finn példái (*kaappaan* és *kääppään*) 'elfog' és 'hurok' jelentésűek, tehát mindkettő 'elkap' értelmű. Bokruknak a finn szövegben nem találtuk nyomát.

A CzF. a „kapa" szót is a *kap(ar)* gyökből származtatja, nem hallgatva el szláv kapcsolatait sem. A TESz. viszont *kapa* szavunkat már egyenesen a szlávból eredezteti, és ennek egyik példájaként a szlovák „kopat" (ásni) szót adja meg. A *kapa, kapál, kapar* szavakat ugyan a vizsgált szövegrész nem tartalmazza, de próbaképpen hozzácsatolhatjuk ezeket is a szlovák és a magyar szöveghez – lássuk, mekkora bokrok ágain ülnek. A magyar eredményt már láttuk: a *kapa, kapál* szavak szervesen illeszkednek a *kap* gyök bokrába. A szlovák találatok száma ehhez képest igen kicsi, mind az ásáshoz és az ásandó dombhoz (kiásott kupachoz?) kötődő jelentéssel. A két bokor összehasonlítása ezért nemcsak méretük, de fogalomgazdagságuk alapján is sokkal

inkább a CzF.-et igazolja, és az átvétel fordított irányát valószínűsíti, ezt pedig a török és a magyar bokrok nagyfokú fogalmi átfedése csak tovább erősíti.

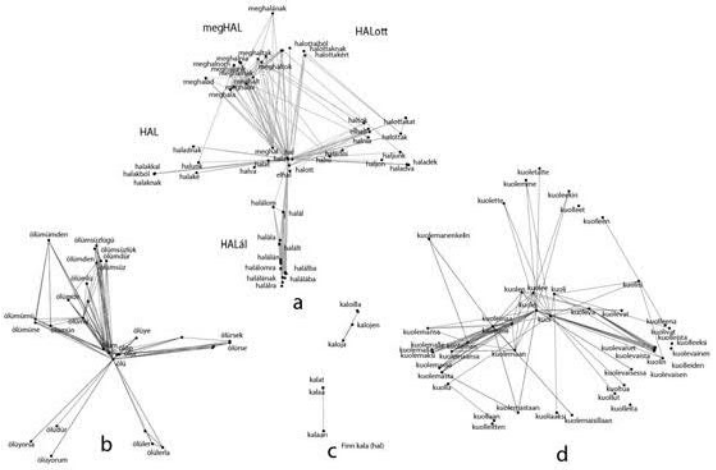


7. ábra

A magyar „kap" (a), a török „kap" (tároló, b) és a szlovák „kopat" (ásni, c) bokrai a megfelelő szövegekben.

Az előbbihez hasonló kérdés vethető fel „hal" gyökünk kapcsán. A 8. ábra ui. a hallal és a halállal kapcsolatos szavakat egy közös bokorban mutatja, ezzel szemben mind a CzF., mind a TESz. két külön szóról beszél. A hal finn megfelelője, a „kala" csak akkora bokrot mutat, amekkora a magyar bokor „halas" nyalábja. Érdekes azonban, hogy mindkét szótár megemlíti a *halál* finn megfelelőjét („kuola"), ám egyik sem utal a „kala" és a „kuole" alaki hasonlóságára. Annál is érdekesebb ez, mert a finn – magyar internetes szótár még a „kalma" és „kuol" (halál) alakot is ismeri. A „kuola" bokra igencsak méretesnek bizonyult – igaz, magát a feltételezett gyököt („kuol") hozzá kellett adni a szöveghez, mert önállóan nincs benne. (A kiegészítés nélkül is nagy, csak éppen töredezett bokrot kapunk.) A „kala" és „kuol" bokrain már jól látszik, hogy ha ezeket egyesítenénk, a magyarhoz nagyon is hasonló eredményre jutnánk. Figyelemre méltó az is, hogy

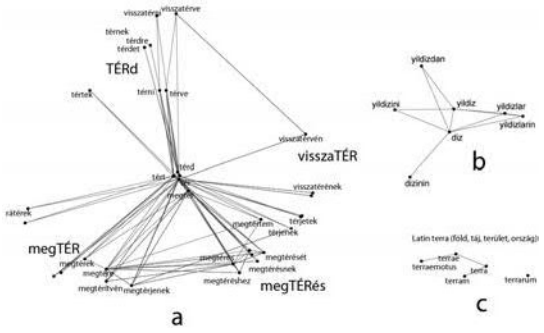
a CzF. a szerinte néma *h*-val kezdődő török „'ölüm' (*ölü* – halott; *ölüm* – halál; *ölüyorsa* – haldokló…; *ölümsüz* – halhatatlan…, stb.) és „balik" (hal) szavakat is a magyar megfelelők rokonának tartja, és bár ezeket módszerünk a merev hangzó-megfeleltetés miatt jelenleg még éppúgy nem képes közös bokorban „érzékelni", mint a finn „kuola-kala" párost, azért a magyar – finn – török egyezés mindenképp említést érdemel[8]. A *hal-halál* kapcsolatról a magyar szájhagyomány is tud – írja már Ipolyi Arnold a *Magyar Mythologia*-ban (Ipolyi, 368-373. o.). A következő önöntés-szöveg pedig, mely a betegséget „halasztja" meg, Takács György még kiadatlan székelyföldi gyűjtése: *„Kérlek téged, Jézus Krisztus, kínszenvedésedre, / Abban a testben ne adj helyet a betegségnek, / Hanem menjen a halak torkába, / Hol a kalászos nem sütődik, / Ahol a kakasszó nem hallszik, / Ahol semmiféle oktalan állatnak nem árthat!"* (Kászonaltíz).



8. ábra
A magyar „hal" (a), a török „ölü' (halott, b), a finn „kala" (hal, c) és a finn „kuol" (halál, d) bokrai.

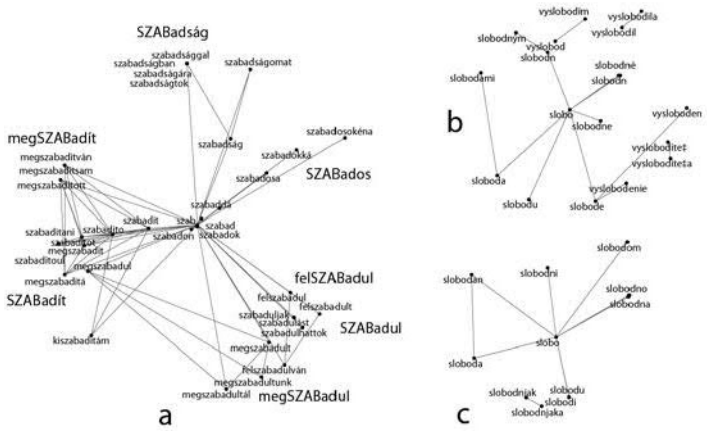
Akár véletlen azonban a „hal" ige és főnév alaki egybeesése, akár nem, a 8. ábra magyar bokrának szerkezete jól mutatja, hogy a jellemző toldalékok elkülönülése miatt módszerünk képes térben szétválasztani a különböző jelentésű szócsoportokat – esetünkben pl. a „halál…" és a „hal…" nyalábjait. A 7. ábrára visszatekintve pedig a „kap" különböző jelentéstartományairól is elmondhatjuk ugyanezt. Ebben az értelemben tehát módszerünk a **homonímiát** is „észreveszi".

Még egy török kapcsolatú gyökünk vizsgálatára nyílik mód az adott szövegtestek segítségével: a „tér"-re. A CzF. török, perzsa, szanszkrit és latin párhuzamokat sorol, a TESz. ebből a törököket teszi magáévá, olyannyira, hogy azt tartja a magyar szó forrásának. A számos elfordulást, mozgásváltást jelölő példa közül csak a szabályos hangváltással származtatható török „diz" adott találatokat, „térd"[9] és „csillag" jelentésekkel[10]. A két bokor méreteit összevetve elég nehéz arra következtetni, hogy lábtérítő szervünk neve török származású volna. Talán mégis fordítva? Annyi mindenképpen kijelenthető, hogy a „térd" magyarul is megáll a lábán. („Térgyes" – mondja Árvai Sándor palóc pásztor a szándékosan íves kitérésűre görbített pásztorbotra.) A latin „terra" ('terület, föld, vidék') szó rokonságát a CzF. említi: A latin bokor mérete szintén eltörpül a magyar mellett.



9. ábra
A magyar „tér" (a), a török „diz" (térd, b)
és a latin „terra" (föld, terület, c) bokrai.

Szab gyökünk hatalmas bokrát mutatja a 10. ábra. A szövegtestben a szabadságra, szabadulásra utaló szónyalábok vannak túlsúlyban, de azért megvan a „szabály", „szablya", „szabados", „szab(játok)" is. Mivel „szabadság" szavunkat a TESz. szláv eredetűnek tartja, érdemes összevetni a magyar bokrot a szlovák[11] és horvát „slobo" (szabadság) megfelelő bokraival, és megállapítani, hogy szabadság szavunk a „szab"- gyök méreteiben és fogalomgazdagságában is igen jelentős bokrába illesztve szláv rokonítás nélkül is jól értelmezhető.



10. ábra
A magyar „szab", a szlovák „slobo" (szabadság, b)
és a horvát „slobo" (szabadság, c) bokrai.

5. ÖSSZEGZÉS

A gyök-elmélet egyik fontos célja az egy gyökből kifejlődött szóbokrok azonosítása. Ezt a feladatot Czuczor Gergely és Fogarasi János egyszer már úgy elvégezte, hogy az utókor másfél évszázadnyi elhallgatás után

is csodálattal fedezi fel újra szótáruk nagyszerűségét, és talál benne iránymutatást a folytatásra. A dolgozatban ismertetett számítógépes rendszer éppen a folytatást kívánja szolgálni. A rendszer egyik pillére egy szórokonságot számszerűsítő algoritmus, amely alkalmas a gyökalapú rokonság felismerésére is, ám működési elvében nem korlátozódik a CzF. gyökszótárának puszta újragyártására. Egyik fő célunk ui. éppen az volt, hogy bebizonyítsuk: létezik olyan független és ésszerű gépi algoritmus, mely a szavak kapcsolatrendszerét bizonyíthatóan és szemmel láthatóan éppen a gyökelmélet elgondolásának megfelelően képezi le. A szemmel láthatóság itt nem jelkép: a számítógépes rendszer másik pillére ui. éppen egy olyan algoritmus, mely a szavak „távolságmátrixából" megszerkeszt egy azoknak optimálisan megfelelő pontrendszert, ha úgy tetszik: szó-univerzumot, vagy annak egy kiragadott részét, a szóbokrot. A 2. ábra szótérképe valóban több bokorszerű rendeződés „galaxisaként" tárta elénk a vizsgált szövegtest szókészletét. Mivel pedig a normált Levensthein-távolság adta kép ettől gyökeresen elűt, az is nyilvánvalóvá vált, hogy ez a rendeződés éppen szórokonító algoritmusunk sajátságai miatt alakult ki. Bár ebben a dolgozatban nem részleteztük, de az is igazolható, hogy ez az algoritmus a különböző nyelvek „gyök-elvűségének" fokát is képes lehet megmutatni.

A dolgozat második részében azt próbáltuk igazolni, hogy e szóbokrok fürtjei, nyalábjai az esetek többségében jól ábrázolják a szóbokrok értelmileg elkülöníthető részeit is. Ezért sorra vettük a szövegtestünkben gyökként viselkedő háromtagú szavak bokrait (pontosabban azoknak egy részét), és megmutattuk, hogy a különböző jelentések, jelentésárnyalatok valóban a bokrok különböző nyalábjaiban tömörülnek, ha a jelentés eltéréseit minimális alaki különbségek is megjelenítik. És habár a homonim gyököket nem tudjuk szétválasztani, azért az ilyenek nyalábjai az eltérő toldalék-rendszerek miatt úgyszintén elkülönülnek a közös bokorban. Ez az eredmény arra mutat, hogy a toldalékoló nyelvekben mindenképpen van létalapja a szóalak és a tartalom összefüggésének, függetlenül attól, hogy természetesen a kivételek is hosszan sorolhatók.

Majd továbbmerészkedtünk, és nyelvismeret hiányában is megkíséreltük feltárni a magyar gyököknek megfelelő rokon szóbokrokat. Merészségünk azonban odáig nem terjedt, hogy pusztán algoritmusunkban bízva saját rokonításokkal álljunk elő. Ezért csakis a CzF. és a TESz. rokonításait vettük alapul, és az ezek mögött rejlő – vagy nem rejlő – szóbokrokat kerestük. Ehhez meg kellett szerezzük az *Újtestamentum* digitalizált szövegrészeit a szóba jöhető nyelveken. Így ui. valóban „szóról szóra" egymásnak megfelelő, gazdag szövegforrásokhoz jutottunk, ezért a bennük feltárható szóbokrok méretei valóban sokat elárulhatnak a közös gyökök különböző nyelvekbeli „otthonosságáról". Ha ui. egy szó valamely nyelvben egy nagy bokor tagja, akkor nehéz feltételezni, hogy újonc jövevényként is pont beillett oda, ha viszont magányos, akkor sokkal könnyebb arra gondolni, hogy jövevény volta miatt vannak „beilleszkedési nehézségei". Az azonos szövegtest még további összehasonlítási alapot is nyújt: ha egy szó ugyanabban a szövegben az egyik nyelven sokszor, a megfelelője a másikon alig fordul elő, az azt mutatja, hogy a második nyelv az adott jelentést inkább más szavakkal fejezi ki – ez pedig ismét sokat elárul az otthonosságról.

Bár a vizsgált mennyiség nyilván nem reprezentatív, mégis szembeszökő, hogy a gyökként viselkedő szavak rokonai közt a finnugorok mind a CzF., mind a TESz. szerint többségben vannak. Meg kell jegyezni ugyanakkor, hogy a CzF. mongol, kínai, szanszkrit, perzsa, stb. rokonításait szövegtest híján nem vizsgáltuk, de ha ezeket is bevonhatnánk a vizsgálatba, a kép jelentősen módosulhatna.

A finn rokonság bokrai közt találtunk a magyarral összevethetőket mind méretük, mind fogalmi terük tekintetében (*tud, hal* és rokonai), és olyat is, melynek bokra formailag igen, de tartalmilag csak egy nagyon szűk mezsgyén kapcsolódik a magyarhoz (*jel* és rokona). A *hal* és a *halál* kapcsolatát éppen a finn és török megfelelők alaki és fogalmi hasonlóságával, valamint a magyar néphagyományban fellelhető fogalomtársítással igyekeztük igazolni, akár még az általunk is ismert bevett álláspont ellenében is. A török kapcsolatok közt a „kap"

gyök esetében hasonló vakmerőségre vetemedtünk: itt a megfelelő gazdag török bokorral igyekeztünk alátámasztani az „elkap" és a „kapu", sőt még a „kapál" fogalmi rokonságát is.

Hasonló érveléssel még „szerencse" és „szabad" szavainkat soroltuk be „szer", ill. „szab" gyökeink bokrába A török rokonszavak közt pedig még a „térd" megfelelőjéről állapítottuk meg, hogy a vizsgált szövegtestben kimutatható bokra a magyar bokor töredéke csupán. Ezek a vizsgálatok felvetik tehát annak a lehetőségét is, hogy megfelelő nyelvföldrajzi vizsgálatokkal kiegészítve egyes kölcsönzési irányokat újraértékeljünk, más, alternatív magyarázatokat keressünk.

E sorok írója sokat foglalkozott a világ népzenei kultúráinak kapcsolatrendszerével, és arra a következtetésre jutott, hogy a mi **népzenénk** egy nagyon régi zenei alapnyelv egyik legközvetlenebb leszármazottja, ezért nem lehet kielégítően jellemezni, megérteni a folytonos és változatos külső hatások puszta eredőjeként. Furcsa volna, ha ez a nyelvünkre nem vonatkozna (Czakó–Juhász 2010).

Egyelőre azonban még csak néhány gyöknél és azok „dilettáns" – inkább a módszer tudományközi alkalmazását felkínálni szándékozó – elemzési kísérleteinél tartunk. Az algoritmus is komoly fejlesztéseket igényel: nyelvenként figyelembe lehetne venni a hangtörténet tényeit, az esetleges hangbetoldásokat, kihagyásokat, hangsorrend cseréket (metatéziseket), a hangszimbolikából fakadó jelentéstani tanulságokat, stb. A dolgozatban azt is hangsúlyoztuk, hogy a szórokonítás alapját képező szókészlet alkalmassága döntően befolyásolja a pontosságot, ez pedig a szövegtestek méretének növelését kívánja meg. A sok hiányosság ellenére mégis abban bízom, hogy ez a dolgozat talán felkelti a **nyelvész szakma érdeklődését**, elősegíti az együttműködést a gyökkutatásban nyelvészet és természettudomány között, az pedig talán elhozza az egész magyarságtudomány kívánt megújulását is.

KÖSZÖNETNYILVÁNÍTÁS

Köszönöm Pomozi Péternek és Czakó Gábornak a dolgozat egyes nyelvészeti következtetéseinek véleményezését, a vadhajtások nyesegetését, Bülent Simşeknek a török szavak magyarázatát.

JEGYZETEK

1 A „szer" egyébként magyarul is jelent összetalálkozást, összerendezést (*Pusztaszer, összeszerel, szeres* település).

2 Sentähden minä myös ahkeroitsen, että minulla aina olisi loukkaamaton *omatunto* Jumalan ja ihmisten edessä.
Ebben gyakorlom pedig magamat, hogy botránkozás nélkül való *lelkiismeretem* legyen az Isten és emberek előtt mindenkor. (Apcsel 24,16)

3 *Tuntemattomalle* Jumalalle… *Ismeretlen* Istennek (Apcsel 17,23)

4 „En usko. Jos en itse näe *naulanjälkiä* hänen käsissään ja pistä sormeani niihin ja jos en pistä kättäni hänen kylkeensä, minä en usko."
Ha nem látom az ő kezein a *szegek helyeit*, és be nem bocsátom ujjaimat a *szegek helyébe*, és az én kezemet be nem bocsátom az ő oldalába, semmiképen el nem hiszem. (János 20,25)

5 Orada ekşİ şarap dolu bİr *kap* vardİ.
Vala pedig ott egy eczettel teli *edény*. (János 19,29)

6 Onlarî bana veren Babam her şeyden üstündür. Onlarî Baba'nın elİnden *kapmaya* kİmsenİn gücü yetmez.
Az én Atyám, a ki [azokat] adta nékem, nagyobb mindeneknél; és senki sem *ragadhatja* ki [azokat] az én Atyámnak kezéből. (János 10,29.)

7 Davut da şöyle diyor: "Sofraları onlara *tuzak*, *Kapan*, tökez ve ceza olsun."
Dávid is ezt mondja: Legyen az ő asztaluk *tőrré*, *hálóvá*, botránkozássá és megtorlássá. (Rómaiakhoz 11,9)

8 Kun hän sitten oli noussut *kuolleista*, opetuslapset muistivat nämä hänen sanansa, ja he uskoivat kirjoituksiin ja siihen, mitä Jeesus oli puhunut. İsa *ölümden* dirilince öğrencileri bu sözü söylediğini hatırladılar, Kutsal Yazıya ve İsanın söylediği bu söze iman ettiler.
Mikor azért feltámadt a *halálból*, megemlékezének az ő tanítványai, hogy ezt mondta; és hivének az írásnak, és a beszédnek, a melyet Jézus mondott vala. (János 2,22.)

9 Sonra *diz* çökerek yüksek sesle şöyle dedi: "Ya Rab, bu günahı onlara yükleme!"
Térdre esvén pedig, nagy fenszóval kiálta: Uram, ne tulajdonítsd nékik e bűnt! (Apcsel 7,60)

[10] A *yildiz*: talán az égitestek éves (*yil*) körforgására, *visszatérő* mozgására utaló kifejezés? Ez esetben a török fogalmi kör a magyarhoz illeszkedve bővülne.

[11] Poznáte pravdu a pravda vás *oslobodí*.
És megismeritek az igazságot, és az igazság *szabadokká* tesz titeket. (János 8,32)

HIVATKOZÁSOK

Benedetto, D. – Caglioti, E. – Loreto, V. 2002. Langue trees and zipping. *Phys. Rev. Lett.* 88/4. 1–4.

Bencze L. – Csébfalvi K. 2002. Genetika – számítástechnika – tudásszociológia. Az ugor-török háború új szakasza avagy vége? In: Büky László – Forgács Tamás szerk. *A nyelvtörténeti kutatások újabb eredményei III. Magyar és finn-ugor jelentéstörténet.* Szegedi Tudományegyetem, Magyar Nyelvészeti Tanszék, Szeged. 7–13.

Borg, I. – Groenen, P. 2005. *Modern Multidimensional Scaling: theory and applications.* 2nd ed. Springer Verlag, New York.

Cohen, W. – Ravikumar, P. – Fienberg, S. 2003. *A Comparison of String Metrics for Matching Names and Records, Communications* Vol. 3, Publisher: Citeseer. 73–8.

Czakó G. 2008. *Beavatás a magyar észjárásba.* Budapest: CzSimon Kiadó.

Czakó G. – Juhász Z. 2010: *Beljebb a magyar észjárásba.* Budapest: CzSimon Kiadó.

CzF. = Czuczor G. – Fogarasi J. 1862–1874. *A magyar nyelv szótára.* Pest: Emich Gusztáv magyar akadémiai nyomdásznál. A digitális változat: Arcanum DVD könyvtár VI. 2004.

Juhász Z. 2011. Low dimensional visualisation of folk music systems using the self organising cloud. *Proc. of 12th International Society for music information retrieval.* Miami, Florida, USA.

Kiss G. – Kiss M. – Sáfrány-Kovalik B. – Tóth D. 2011. *A Magyar szóelemtár megalkotása és a Magyar gyökszótár előkészítő munkálatai,* In: VIII. Magyar Számítógépes Nyelvészeti Konferencia MSZNY 2011. Szerk.: Tanács Attila, Vincze Veronika, Kohonen, T. – Somervuo, P. 1998. Self-organizing maps of symbol strings. *Neurocomputing* 21. 19–30. Szeged.

Laki L. – Prószéky G. 2010. Statisztikai és hibrid módszerek párhuzamos korpuszok feldolgozására. In: Tanács A. – Vincze V. (szerk.) *A VII. Magyar Számítógépes Nyelvészeti Konferencia előadásai.* Szeged: Szegedi Tudományegyetem, 9–79.

Mc Eney, T. – Wilson, A. 1998. *Corpus Linguistics.* Columbia University Press.

MESz. = Tótfalusi I. 2004. *Magyar etimológiai szótár* Arcanum DVD könyvtár.

Navarro G. 2001. A guided tour to approximate string matching. *ACM Computing Surveys* 33/1. 31–88.

Terra, E. – Clarke C. L. A. 2003. Frequency estimates for statistical word similarity measures. *Proceedings of the 2003. Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology.* Vol. 1. Association for Computational Linguistics Stroudsburg, PA, USA.

TESz. = *A magyar nyelv történeti-etimológiai szótára* 1–3. Főszerk. Benkő Loránd. 1967, 1970, 1976. Budapest: Akadémiai Kiadó.

A BIBLIAFORDÍTÁSOK DIGITALIZÁLT SZÖVEGEINEK WEBOLDALAI:

Magyar: http://www.sacred-texts.com/bib/wb/hun/index.htm Károli Gáspár 1590.
Finn: http://www.sacred-texts.com/bib/wb/fin/index.htm 1776. (1859.)
Török: http://www.sacred-texts.com/bib/wb/trk/index.htm
Horvát: http://www.sacred-texts.com/bib/wb/cro/index.htm
Szlovák: http://www.biblegateway.com/versions/index.php?action=getVersionInfo&vid=40#booklist, 1993.