

ENDRÉDY ISTVÁN – PRÓSZÉKY GÁBOR

A Pázmány Korpusz

Pázmány Corpus, the largest Hungarian annotated corpus with 1.2 billion tokens has been created mainly to support the ongoing computational psycholinguistic research of our language technology group. The big text database collected automatically from the web was first processed by a special boilerplate removal algorithm then categorized according to their distance from the texts written according to the academic orthography. The corpus is then annotated on various levels: lemmas, various POS tags (made by different POS taggers) and noun phrase labels.

Keywords: corpus, automatic corpus building, boilerplate removal, Humor, PurePos, HunTag3, annotated corpus

1. Magyar nyelvű szövegtörzsek

Az elérhető magyar korpuszok száma öröndetes módon egyre növekszik. A mintavétel vagy a felhasználás módjától függően nagyon sokféle korpusz létezik: vannak egynyelvű és többnyelvű korpuszok, általánosak és szaknyelvi, statikusak és dinamikusak, szinkrónok és történetiek. Jelen írásunkban csak az egynyelvű – ezen belül is csak a magyar nyelvű – szinkrón korpuszokkal foglalkozunk. Az egyik legnagyobb ezek közül a 2004-ben a BME MOKK által készített webkorpusz (HALÁCSY ET AL. 2004), melynek mérete 589 millió szó, ám ezek mindenféle nyelvészeti annotáció nélküli „tisztá” szövegek. A Magyar Nemzeti Szövegtár a maga 187,5 millió szavas méretével ugyan ennél kisebb, ám komoly előnye a tudatosan válogatott tartalom, és az, hogy annotált, azaz a teljes szöveg minden szavához szófaji információ is tartozik, amit egy automatikus szófaji egyértelműsítő program kiegészítette a szöveg szavait az aktuális szövegbeli pozícióban a szó tövének helyes szófajával (VÁRADI 2002). Ennek a korpusznak már készül az újabb változata, az MNSZ2, mely folyamatosan növekszik (ORAVECZ ET AL. 2014): mérete jelenleg 1,04 milliárd tokenes.

A mai magyar nyelvtechnológiai kutatások legrégebb óta széles körben használt korpusza az 1,2 millió szavas Szeged Korpusz (ALEXIN ET AL. 2003). Ez ugyan lényegesen kisebb az előzőknél, de egyedülálló abban, hogy teljes morfológiai annotációja emberi ellenőrzéssel készült. Ezt egyébként a gigakorpuszokon gyakorlatilag nem is lehet megvalósítani, épp a hatalmas méretből adódó jelentős időigény miatt. Feltétlenül meg kell még említeni a Szeged Korpusznak az egyes mondatrészekre vonatkozó annotációval kiegészített változatát, melynek neve: Szeged TreeBank (CSENDES ET AL. 2005).

Pszicholingvisztikai indíttatású magyar nyelv kutatásunk kezdetén, tehát a 2010-es évek elején még nem volt elérhető olyan igazán nagy méretű (tehát milliárd tokenes) szövegtörzs, amelyik többféle annotációjával a performancia-alapú számítógépes elemzésre irányuló kutatási projektünk (PRÓSZÉKY – INDIG 2015) igényeit megfelelően ki tudta volna szolgálni. Így 2012-ben kifejlesztettünk egy ún. crawlert, azaz az internet magyar szöveges tartalmainak számunkra megfelelő sebességű és minőségű automatikus letöltésére szolgáló speciális programot (ENDRÉDY – NOVÁK 2013). Ennek segítségével a weben található magyar szövegekből minél többet szándékoztuk összegyűjteni. Ez a gyakorlatilag állandóan futó szoftvereszköz évek alatt összegyűjtötte a korpusz alapanyagát, amivel meg tudtuk kezdeni a korpusz nyelvi tartalmának összeállítását.

2. A korpusz alapjául szolgáló weboldalak összegyűjtése

Az írott nyelvi szerkezetek kutatását célzó automatikusan épített korpuszoknál a szövegek szószámában megadott mérete mellett igen fontos szempont a korpusz minősége. Egy frissen építendő korpusz esetében a minimális elvárás a duplikátummentesség, valamint a minél tisztább – azaz az interneten oly gyakran megjelenő, de a törzsanyaghoz nem tartozó reklámoktól mentes – szövegekből álló adatbázis. Az e célra szolgáló módszerek közül (BIEMANN ET AL. 2013) a következők szempontok szem előtt tartásával igyekeztünk tartani a minőséget: figyeltük a korpusz n leggyakoribb szavát, a legrövidebb és a leghosszabb mondatok hosszát, illetve a leggyakoribb kategóriasorozatokat (ez utóbbira a 3. táblázatban található példák).

A legtöbb web-alapú szövegtörzs esetében először nagy mennyiségben töltenek le HTML-lapokat az internetről, majd ezekből különféle automatikus technikákkal kinyerik a szöveget, végül egy utófeldolgozó eljárással megtisztítják a duplikátumoktól az adatbázist. Ez utóbbi két lépés különösen kritikus a végeredményként előálló korpusz minősége szempontjából. A szövegekinyeréskor fellépő tipikus problémák egyike, hogy ha bent marad zajos szöveg, akkor az a későbbi feldolgozást is nehezíti, az eredményeket torzítja, hiszen valójában oda nem való, nem kívánt részeket (reklám, menüsor stb.) is tartalmaz. Az internetes oldalakon oly gyakori menüsorok megfelelő strukturált szöveggé¹ alakítása minden mondatszegmentálót² nehéz feladat elé állítanak. A duplikátummentesítés pedig egy olyan, gépi eszközök által irányított folyamat, melyben a folyó szövegből kitörölődnek bizonyos részek, amiket a program duplikátumnak minő-

¹ A strukturált szöveg olyan annotációval ellátott szöveg, melyben jelölve vannak a mondathatárok, a szavak szófaja, szótöve, vagy kiolvasható a szöveg szerkezetére utaló tetszőleges más metainformáció.

² A mondatszegmentáló az a modul, ami a folyó szöveget mondatokra bontja.

sít, így az eredményül kapott szöveg nem feltétlen lesz összefüggő, koherens szöveg, azaz a korpusz mondatai néhol nem vagy nem megfelelően kapcsolódnak egymáshoz. Mivel kutatásunk a többmondatos megnyilvánulásoknak az emberéhez hasonló gépi kezelését (aminek legjobb példája a sajtóban megjelenő rövidhírek) célozza meg, ahol tehát néhány mondatos szövegegységek értelmezése a cél, nagyon fontos, hogy ne csak az egyes mondatok, hanem az összefüggő mondat-*n*-esek is együtt maradjanak.

Az általános, illetve a speciális webkorpuszok építésének különbségei régóta szolgáltatnak kutatási témát a szakembereknek (BARONI – UEYAMA 2006). A mai világban a marketingcélokat jól támogató webes kommenteket sokszor gyűjtik ki speciális korpuszokba véleménybányászat (opinion mining) céljából, például ilyen a Birmingham Blog Corpus vagy a CROW (NEUNERDT ET AL. 2011). Magyar nyelvre is készült ilyen kommentkorpusz (MIHÁLTZ ET AL. 2015): ez 226 ezer felhasználó 1,9 millió Facebook-hozzászólását vizsgálta alapvetően politikai szempontból.

Általános korpuszunkhoz először is nagy mennyiségű letöltött weblapra volt szükségünk, ehhez egy egyszerű letöltő eszközt, ún. crawlert készítettünk. A web-oldalakon található szövegek kinyerése azt jelenti, hogy a beolvasott HTML-tartalomnak csak egy töredéke, mintegy 15-20%-a az, ami a nyelvi feldolgozás számára az igazi nyersanyag, azaz valódi magyar nyelvi szöveg. Eleinte az eredeti HTML-tartalmakat is tároltuk, így a későbbi procedurális módosításoknál, hangolásoknál az utómunkák során az újrafuttatás nagyságrenddel gyorsabban zajlott, hiszen a tevékenység nem igényelt újabb letöltéseket, viszont ennek az eljárásnak jelentős a tárigénye. Az így előállt szöveges adatbázist további tisztításnak és utómunkáknak vetettük alá. A magyar lapok karakterkódolásukban ugyanis még ma sem egységesek, mert többféle kódlap van használatban: elsősorban az iso-8859-2, utf-8, iso-8859-1, de az is gyakori, hogy a magyar ékezetes betűk HTML-kódolással szerepelnek. A szövegeket ezért egységesen a sztenderd utf-8 kódlapra konvertáltuk.

3. Automatikus szövegkinyerés a weboldalakból

A webről nyert korpuszok építésekor az általában dinamikusan generált HTML-tartalomból történő szövegkinyerés, mint említettük, nem triviális feladat a különböző oldalakon ismétlődő rengeteg irreleváns sablonos tartalom miatt. Ennek a technikai lépésnek, mely kiemeli a később használandó nyelvi tartalmat a rengeteg egyéb, weblap-specifikus információ közül, az angol irodalomban boilerplate removal,³ azaz sablonszűrés a neve. A művelet lényege, hogy a HTML-

³ Magyarul leginkább sablonszűrőnek lehetne fordítani.

tartalomból csak az értékes részeket igyekszik megtartani, a menük, a fej- és láblécek, a reklámok, a minden oldalon ismétlődő struktúra kiszűrésével (1. ábra).

1. ábra. Boilerplate-algoritmus: az elemzési céllal összegyűjtendő szöveg kiemelése egy weboldalról



A fenti feladatra számos algoritmus létezik. Ezek azon az alapon működnek, hogy a HTML-ben előforduló szövegek bizonyos tulajdonságai alapján dobnak ki (vagy tartanak meg) egy szöveget. Ilyen tulajdonságok lehetnek például a szöveg hossza, a benne előforduló linkek sűrűsége, a HTML-címkék sűrűsége, vagy például az ún. stopwordok⁴ aránya az összes szóhoz képest. Több sablonszűrő megoldást megvizsgáltunk, és az elérhetők között a JusText (POMIKÁLEK 2011) hozta a legjobb eredményt. Azonban ez is tévesztett bizonyos web-oldalakra: a kommenteket sokszor a cikkhez vette, a vizsgált tulajdonságok alapján ezen részeket is „cikknek” minősítette. Ezért egy új megoldás fejlesztése mellett döntöttünk.

A problémát az adott egyedi weboldalra magasabb szintre lépve sikerült megoldani a korábbinál hatékonyabban. A korábbi boilerplate-algoritmusok elsősorban arra törekedtek, hogy az egyes weblapok nem kívánt részeit távolítsák el. Mikor ezzel nem sikerült átütő eredményt elérni, akkor döntöttünk úgy, hogy pozitív gondolkodásra váltunk, és az értékes részeket fogjuk keresni. Így született meg az Aranyásó algoritmus (ENDRÉDY – NOVÁK 2013, ENDRÉDY 2016). A nevét onnan kapta, hogy egy aranyásó több köbméter salakot is átszítál egy picit aranyrögért, és ha megvan, örül neki: ez az algoritmus sok lapot átnéz egy mintázat után kutatva, amit aztán megtanulva az adott webtartomány többi lapjaira is alkalmaz.

⁴ Stopword: gyakran előforduló, általában rövid funkciószavak, melyeket a keresési szempontok alapján optimalizált szövegtárolásnál nemigen szokás figyelembe venni. Leggyakrabban a névelők, a kötőszavak, a módosítószók, illetve egyes nyelvekben a prepozíciók is ilyenek.

A megoldás azon alapul, hogy egy ún. doménon azaz web-tartományon belül a dinamikusan generált weboldalak, illetve url-ek nem szöveges tartalma (pl. a HTML-kód) általában tartalmaz közös mintázatokat, hasonló dizájnelemeket, amelyeket azonosítva megtalálható a köztük rejlő értékes tartalom. Az algoritmus az adott domén oldalairól mintát vesz, és megpróbálja megkeresni azt a HTML-címkét, amelyen belül (az oldalak zömében) a cikk található, különös tekintettel azokra a mintákra, amelyek az oldalakon ismétlődnek. Például gyakori a hírportálokon, hogy a cikk alján feltüntetik a legnépszerűbb öt cikk ajánlóját. Ezeket a szövegeket a korábbi sablonszűrő algoritmusok nem szűrik ki (pl. a JusText a cikk részének tekinti őket), mert önmagukban ezek jó minőségű szövegek, ám duplikátumot okoznak majd a korpuszban, amivel erősen felülreprezentálják a bennük rejlő tartalmakat.

Az Aranyásó algoritmus egyfajta előtétként működik a JusText előtt: csak azt a HTML-kódot adja át neki, ami nagy eséllyel az értékes részt tartalmazza, a többit már eleve eldobja, így az őt követő sablonszűrőnek jelentősen leegyszerűsíti a feladatát. Más szavakkal: minden doménre megtanuljuk azt a HTML szülőcímkét, amely csak a cikket tartalmazza, majd csak ennek a tartalmát adjuk oda a JusText sablonszűrő algoritmusnak. Ennek a módszernek az az előnye, hogy azok az oldalak, ahol nincs cikk (tematikus nyitólapok, címkefelhők, keresőlap eredmények stb.), ott az algoritmus nem ad semmit, hiszen a doménre jellemző cíkcímke nincs jelen. Így az algoritmus automatikusan kiszűri ezeknek a lapoknak a tartalmát, azokon a lapokon pedig, ahol valóban van cikk, a sablonszűrő algoritmus már csak a lényegi tartalmat kapja. Crawlerünkben tehát főként az Aranyásó algoritmus végezte a HTML-tartalomról az értékes szöveg kinyerését, ami a JusText megoldásra épül, és azt az adott domén felépítésére jellemző ismerettel bővíti ki, ezzel is javítva a pontosságot. (Endrédi 2016)

4. Algoritmikus osztályozás: összefüggő szöveg és felsorolások, illetve sztenderd szövegek és kommentek

Az összegyűjtött korpusz szövegei meglehetősen eltérő jellegűek voltak: egyrészt cikkek, blogok, elbeszélések, azaz többé-kevésbé folytonos olyan szövegek, melyeket javarészt a sztenderd helyesírási ajánlások betartásával írtak, másrészt termékfelsorolások, címkefelhők, vagy a leggyakrabban keresett kifejezések, amik nem igazán folyamatos szövegszerű anyagok, bár ezek általában a helyesírási normától nem nagyon térnek el. Az összefüggő szövegek és a felsorolások viszont vegyesen fordultak elő, ezért szükségesnek tartottuk, hogy lehetőség szerint ezeket algoritmikusan elkülönítsük egymástól. Az egész korpuszt ezért átszűrtük a következő módon: ha egy web-lap bekezdéseinek átlagos hossza és a stopwordök aránya egy adott küszöbérték alatt volt, akkor az oldal az „Egyéb tartalmak” kategóriába került (1. táblázat). Egyébként megtartottuk ezt a

szöveghalmazt is egy esetleges későbbi feldolgozás céljából, bár ezek nem olyan szövegek, amiket „olvasni szoktunk” (pl. felsorolások). Felosztásunkkal sikerült elérni, hogy a korpusz „sztenderd szövegek” nevű részébe csak olyan szövegek kerültek, melyeket valószínűleg átolvastak, mielőtt a webre tették volna őket.

1. táblázat: A Pázmány Korpusz összetétele

alkorpusz	tokenszám	főnévicsoportszám	mondatszám
sztenderd szövegek	954 424 307	223 347 534	48 536 849
egyéb tartalmak	228 920 436	52 865 889	15 802 499
kommentek	59 013 104	13 867 066	3 505 818
összesen	1 242 357 847	290 080 489	67 845 166

A webes korpusz építése során felmerült az ötlet, hogy a felhasználók által készített hozzászólások szövegeit érdemes volna nyelvi szempontból külön is megvizsgálni. Ezek a tartalmak önmagukban nagyon eltérő minőségűek, hiszen nem újságíróktól és szerkesztőktől származnak, hanem szabadon hozza létre ezeket bárki, olykor keresetlen szavakkal és sok érzellemmel. Emiatt úgy döntöttünk, hogy létrehozunk számukra egy külön alkorpuszt, a kommentkorpuszt. A leendő felhasználók számára hasznos lehet a minél több típusú szöveg, és így az elkülönítéssel egy adott kutatásban az alkorpuszok akár külön-külön is vizsgálhatók lesznek. A nagyobb hírportálokról összegyűjtött hozzászólásokból végül egy 59 millió szavas alkorpusz keletkezett. Ehhez annak az algoritmikus meghatározására volt szükség, hogy egy szövegben hol ér véget egy cikk, és hol kezdődnek a kommentek. Kommentnek minősítettük azokat a szövegeket, amelyekben bizonyos erre a szövegtípusra jellemző mintázatok fordultak elő, mint például: keresztnév dátummal vagy számmal, dátum #123, <szám> hozzászólás, „Hozzászólások 123”, „nickname 2014.02.02.”, stb. Az is sokszor segített, hogy ezeknek a szövegeknek egyik – a törzsszövegektől eltérő – jellemzőjük, hogy sűrűn fordulnak elő bennük emotikonok.

5. A korpuszannotáció lépései

A korpuszok kutatási használhatóságát nagyban növeli, ha többféle annotáció is szerepel bennük. Mivel ilyen mennyiségű szövegen kézi annotációval előállított precíz szöveg (az ún. gold standard) nem hozható létre,⁵ ezért gépi eljárásokkal a

⁵ Becslések szerint az összegyűjtött szöveganyagnak a folyamatosan történő olvasása önmagában mintegy 90 évet venne igénybe.

következő annotációkkal láttuk el a szöveget: mondathatárok, lemma, szófaj, valamint a főnévi csoportok kezdete és vége.

A Huntoken modul (HALÁCSY ET AL. 2004) volt a feldolgozás során az első nyelvfeldolgozó modul: ezzel mondatokra bontottuk a szöveget. Bár a duplikátummentességre már a HTML szövegkinyerésénél is törekedtünk, ám hiába unikus minden letöltött URL és a kinyert tartalom is, még így is előfordult, hogy szinte ugyanaz a tartalom többször is letöltésre került. Ez egyrészt annak volt köszönhető, hogy a portálokon a cikkek egy idő után archívumokba kerülnek, azaz kapnak egy másik URL-t, vagy egy cikk egyszerűen több URL-en is elérhető (pl. belfold.domain.hu/cikk123 és hirek.domain.hu/cikk123). Másrészt előfordul, hogy a tartalomban valami egészen apró módosítás történik, aminek következtében a megváltozott szöveg a gépi feldolgozás számára értelemszerűen újnak fog számítani. Ezért egyrészt nagy figyelmet fordítottunk arra, hogy már a szövegkinyerésénél csak a lényeges tartalmat vegyük ki, ugyanis

(1) ez csökkenti a fő tartalom körüli, ismétlődő szövegek megjelenését a korpuszban,

(2) és így szigorúbban lehet szűrni a cikk szövegének egyediségére (ha egy cikk már szerepelt, akkor nem vesszük fel), másrészt – a biztonság kedvéért – a végső korpuszösszeállításnál duplikátummentesítést végeztünk nemcsak URL- és cikk-, hanem bekezdés- és mondat szinten is: ez utóbbit már a Huntoken kiemenetén.

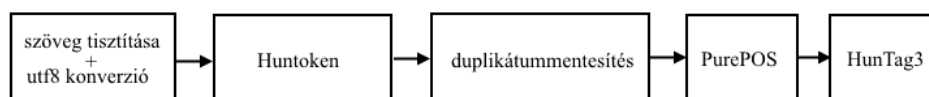
A lemmákat a Humor morfológiai rendszerre (PRÓSZÉKY – TIHANYI 1993) épített lemmatizálóval⁶ (ENDRÉDY – NOVÁK 2015), a szófaji egyértelműsítést pedig az erre a lemmatizálóra is építő PurePOS szoftvermodul (OROSZ – NOVÁK 2013) alkalmazásával végeztük. Ezen túl egy új, pontosabb szófaji kódolással (LIGETI-NAGY 2015a) is annotáltunk, mely a korábban használtaknál részletesebb, finomabb kategóriákba sorolja a szavakat. Például a főneveknél külön címkét kapnak a tulajdonnevek, a foglalkozások, a napok, a hónapok, az anyagnevek stb. Ezt a felosztást az indokolta, hogy részletesebb kategóriacímkék mellett a tanulóalgoritmusok könnyebben tudnak rátanulni egy-egy nyelvi jelenségre. Például a pontosabb <NPROP ADJ NOCCUP> séma segítségével egyetlen szerkezetként lehet a teljes anyagban kezelni az olyan főnévi csoportokat, mint pl. *Obama amerikai elnök*, *Merkel német kancellár*, ám a hagyományos <N ADJ N> minta a *cumisüveg potenciális veszélyforrás* szóhármassal esetén is találatot adna (ENDRÉDY – NOVÁK 2012).

Mint már jeleztük, nemcsak a szófaj és a lemma alapján lehet keresni a korpuszban, hanem teljes főnévi csoportok is kereshetővé váltak a magyar weben

⁶ A lemmatizáló a lemmát, azaz a szóalak grammatikai szótövét állapítja meg.

előforduló szövegekben. A korpuszban ehhez el kellett végezni a főnévi csoportok bejelölését is: ezt a feladatot a *HunTag3* (ENDRÉDY – INDIG 2015) segítségével oldottuk meg. A szövegfeldolgozás lépései az 2. ábrán láthatók.

2. ábra. A szöveg feldolgozásának lépései



Egyetlen mondatnyi korpuszrészlet bemutatásával a 3. ábrán igyekszünk érzékeltetni a Pázmány Korpusz így kialakított szerkezetét, melyben az öt oszlop tartalma: a szóalak, a lemma, a (korábban említett kétféle) szófaji kódolás és a (B-től E-ig tartó, illetve az egyeleműeknél egyetlen egyesből álló) főnévcsoport-jelölés.

3. ábra. A Pázmány Korpusz szerkezete

eredeti szóalak	szótő	Humor- szófajcímke	módosított Humor- szófajcímke	főnévcsoport- határok
<s>				
Kezdetkor	kezdet	N+TEM	FN+TEM	1-N_1
teremtette	teremt	V+TM _{e3}	IGE+TM _{e3}	O
Isten	isten	N+NOM	FN+NOM	1-N_1
az	az	DET	DET	B-N_1
eget	ég	N+ACC	FN+ACC	E-N_1
és	és	CON	KOT	O
a	a	DET	DET	B-N_1
földet	föld	N+ACC	FN+ACC	E-N_1
.	.	PUNCT PERIOD		PUNCT O
</s>				

A korpuszt a Bonito korpuszkezelő rendszer (RYCHLÝ 2007) által használt, az XML-szabványhoz közel álló formátumban tároljuk. A főnévi csoportok határait a 3. ábrán is látható, az utolsó oszlopban jelzett reprezentáció mutatja. A jelölést címkézési feladatokra (sequential tagging) hozták létre, melynek lényege, hogy adott címkékkal jelöljük egy sorozat első (B, *begin*), közbenső (I, *in*), végső (E, *end*) vagy nem beletartozó (O, *out*) elemét is. Többféle reprezentáció létezik (TJONG KIM SANG – VEENSTRA 1999), attól függően, hogy a négy címkéből melyeket használjuk. Korpuszunkban azt a reprezentációt használtuk, amelyet a magyar nyelvre először a Szeged TreeBank annotációjában alkalmaztak (RECSKI 2014), és a legjobb eredményeket produkálta statisztikai tanulóalgoritmuskok esetén. Ez a reprezentáció explicit módon jelöli a sorozat végét, illetve

külön címkét használ az egy ($I-N_1$), a kettő ($B-N_1$, $E-N_1$), és a kettőnél több szóból álló ($B-N_{2+}$, $E-N_{2+}$, $E-N_{2+}$) szerkezetekre.

A részletes morfoszintaktikai elemzés tehát tövekre, toldalékokra, illetve összetételi tagokra bontja a szöveg szavait, továbbá szófaji egyértelműsítést végez, megadva a szóalak tövének – aktuális pozícióbeli – szófaját. A 3. ábrában kódolt főnévcsoport-szerkezet pedig az IOB-kódok értelmezésével a következő:

$$[_{NP} \textit{Kezdetkor}]^{\text{TEM}} \textit{teremtette} [_{NP} \textit{Isten}]^{\text{NOM}} [_{NP} \textit{az eget}]^{\text{ACC}} \textit{és} [_{NP} \textit{a földet}]^{\text{ACC}} \\ [_{\text{PUNCT.}}]$$

Ebből pedig a későbbiekben a főnévi csoportok kiemelésével kapott ún. mondatváz a fenti példában a *teremt* ige egy konkrét vonzatkeret-kiosztását adja:

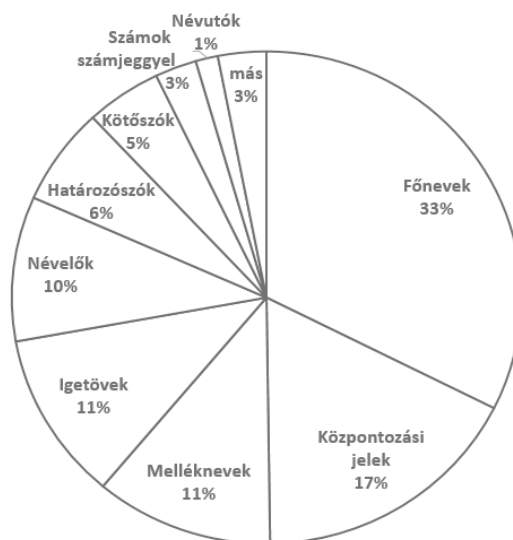
$$[_{NP}]^{\text{TEM}} \textit{teremtette} [_{NP}]^{\text{NOM}} [_{NP}]^{\text{ACC}} [_{\text{PUNCT.}}]$$

6. Eredmények

Az elkészült korpusz szigorú értelemben véve nem kiegyensúlyozott korpusz, hiszen nem nyelvészek válogatták a tartalmát, hanem egy robot bejárásai döntései: kigyűjtött linkeken elindulva heterogén, sokféle jellegű szöveget (újságcikkeket, blogokat, bulvároldalakat, játékportálokat stb.). gyűjtött le az általunk készített szoftverrendszer Szófajkódokkal, lemmával és a főnévi csoportok kezdő- és végcímkeivel is annotáltuk a jelenleg legnagyobb magyar korpuszt (1. táblázat). Mérete épp az ezerszerese a számos kutatás alapjául szolgáló – ám emberi közreműködéssel ellenőrzött, ezért mindenütt pontos annotációkat tartalmazó – Szeged Korpusznak (CSENDES – CSIRIK – GYIMÓTHY 2004, CSENDES ET AL. 2005). A Pázmány Korpuszban szereplő szavak szófaji megoszlását mutatja – alkorpuszok szerinti megoszlásban is – a 2. táblázat, illetve a 3. ábra. A főnévi csoportok szerinti annotációk a magyar névszói szerkezetek felépítésének valós képét vetítik elénk a weben olvasható szövegek alapján. Ezeknek a főnévi csoportoknak az annotálásával, majd a mondatokból egy önálló adatbázisba való át-emelésével megindultak a magyar főnévi csoportok és a kiemelésükkel létrejött mondatváz-szerkezetek gyakorisági vizsgálatát célzó kutatások is (LIGETI-NAGY 2015b). A korpuszban történő keresések támogatására a népszerű *Sketch Engine* (KILGARRIFF ET AL. 2014) korpuszkereső limitált változatát, a *NoSketchEngine* nevű nyílt forráskódú korpuszkezelő programcsomagot használtuk (RYCHLÝ 2007), mely a Bonito nevű grafikus felületből és a mögötte meghúzódó *Manatee* korpuszkezelő eszközökből áll, mely akár több milliárd szavas adatbázist is képes kezelni (4. ábra). Ez a programcsomag már az Magyar Nemzeti Szövegtár újabb változatai esetében is jól bevált gyors válaszidejével. Az eddigieknél morfológiailag pontosabban címkézett (LIGETI-NAGY 2015a) korpuszban a legfrissebb magyar főnévcsoport-annotáló megoldással (ENDRÉDY – INDIG 2015) jelöltük a

főnévi csoportok határait is. Reményünk szerint mindezek már most hasznos lehetőségeket biztosítanak a mai írott magyar nyelvet kutatók számára.

4. ábra. A Pázmány Korpusz tokeneinek szófaji megoszlása



2. táblázat. Szófaji megoszlás a Pázmány Korpuszban

Szófaj	Sztenderd szövegek	Egyéb tartalmak	Kommentek	Összesen
Igék	113 808 282	16 019 286	7 430 599	137 258 167
Finit alakok	88 399 403	11 865 941	5 844 756	106 110 100
Folyamatos melléknévi igenevek	7 264 051	1 453 995	345 309	9 063 355
Befejezett melléknévi igenevek	6 713 811	1 087 470	316 864	8 118 145
Infinitívuszok	8 013 350	1 063 585	714 782	9 791 717
Ragos infinitívuszok	1 091 261	111 631	72 551	1 275 443
Határozói igenevek	2 326 406	436 664	136 337	2 899 407
Főnevek	295 939 929	89 893 104	16 938 977	402 772 010
Tartalmas főnevek	263 185 702	86 120 048	14 201 657	363 507 407
Főnévi névmások	32 754 227	3 773 056	2 737 320	39 264 603
Melléknevek	109 080 322	25 090 697	6 532 678	140 703 697
Alapfokú melléknevek	97 396 981	23 688 145	5 840 962	126 926 088

Középfokú melléknevek	2 483 182	321 407	144 837	2 949 426
Felsőfokú melléknevek	2 436 902	404 614	119 915	2 961 431
Melléknévi névmások	6 763 257	676 531	426 964	7 866 752
Számok számjeggyel	16 028 228	15 702 037	1 154 059	32 884 324
Arab számok	15 931 101	15 637 461	1 151 944	32 720 506
Római számok	97 127	64 576	2 115	163 818
Számnevek	12 309 435	1 480 423	683 171	14 473 029
Valódi számnevek	11 562 409	1 383 794	602 876	13 549 079
Számnévi névmások	747 026	96 629	80 295	923 950
Határozószók	66 641 046	7 759 986	5 172 711	79 573 743
Valódi határozószók	54 417 992	6 422 872	4 144 396	64 985 260
Határozószói névmások	12 223 054	1 337 114	1 028 315	14 588 483
Névelők	99 927 941	12 769 812	5 433 479	118 131 232
Névutók	15 120 529	1 730 167	732 549	17 583 245
Kötőszók	50 376 009	6 073 741	3 417 733	59 867 483
Igekötők	12 266 552	1 484 210	778 154	14 528 916
Egyéb szófajok	3 169 949	1 447 279	270 225	4 887 453
Partikula (-e)	434 889	47 416	24 538	506 843
Központozási jelek	157 263 951	47 979 375	10 314 049	215 557 375
Kötőjelek	54 402	85 221	12 833	152 456
Perjelek (törtvonalak)	211 423	448 798	17 692	677 913
Ismeretlen szavak	1 791 420	908 884	99 657	2 799 961
Összes token	954 424 307	228 920 436	59 013 104	1 242 357 847

3. táblázat. A legtipikusabb magyar főnévi csoportok szerkezete

	A főnévi csoport felépítése	Korpuszbeli gyakoriság	Példa
1	[N+Nom]	14 312 895	<i>ház</i>
2	[Det] [N+Nom]	11 632 547	<i>a ház</i>
3	[N Pron+Nom]	6 416 317	<i>ez</i>
4	[Adj+Nom]	6 279 294	<i>okos</i>
5	[Det] [N+Acc]	4 616 117	<i>a házat</i>

6	[N Pron+Acc]	4 206 866	<i>ezt</i>
7	[N+Acc]	3 639 354	<i>házat</i>
8	[Det] [Adj] [N+Nom]	3 389 714	<i>a nagy ház</i>
9	[N+Nom] [N+Nom]	2 920 072	<i>műanyag ház</i>
10	[Det] [N+Pl+Nom]	2 690 635	<i>a házak</i>
11	[Adj] [N+Nom]	2 576 916	<i>nagy ház</i>
12	[Num Digit+Nom]	2 473 950	<i>2</i>
13	[N+Ine]	2 240 616	<i>házban</i>
14	[N Pron]	1 904 343	<i>ez</i>
15	[Det] [N+Ine]	1 855 041	<i>a házban</i>
16	[N+PSe3+Nom]	1 802 458	<i>háza</i>
17	[N+Sup]	1 653 639	<i>házon</i>
18	[Det] [N+Nom] [N+PSe3+Nom]	1 648 933	<i>a kutya háza</i>
19	[N Pron+Dat]	1 360 964	<i>ennek</i>
20	[Det] [Adj] [N+Acc]	1 356 462	<i>a nagy házat</i>
21	[Det] [N+Sup]	1 306 032	<i>a házon</i>
22	[N+Sub]	1 206 749	<i>házra</i>
23	[N+Pl+Nom]	1 180 336	<i>házak</i>
24	[Adj] [N+Acc]	1 125 795	<i>nagy házat</i>
25	[N Pron+Pl+Nom]	1 125 022	<i>ezek</i>
26	[Num Digit+Ine]	1 111 783	<i>1910-ben</i>
27	[Det] [N+Dat]	1 105 306	<i>a háznak</i>
28	[N+Ins]	1 095 360	<i>házzal</i>
29	[Det] [N+Pl+Acc]	1 092 733	<i>a házakat</i>
30	[Num Digit] [N+Nom]	952 247	<i>2 ház</i>

5. ábra. Néhány keresési eredmény a Pázmány Korpuszból

Query vár, FN.* 103,303 (83.2 per million)		
First	Previous	Page 43 of 4,133 Go Next Last
doc#5	normann lovaggal ostromolta meg Galeria várát , bevette , és foglyul ejtette Benedeket	
doc#5	, VII. a Matild birtokában lévő Canossa várába ment , hogy várja a fejleményeket . 1077.	
doc#5	25 : Henrik vezeklő ruhában jelent meg a vár kapuja előtt és 3 napon át kérte a perc.	
doc#5	Pozsony ostrománál nagykorúították , amikor a várát Borisztól visszavítta , IX. 11 : a Boriszt	
doc#5	ostromolták . A háromszög alaprajzú hegyi várból jelentős romok állnak . Fügedi 1977:55	
doc#5	vízhez , földalatti folyosót vágtak Sion várában , mely egy 13 m mély , függőleges aknához	
doc#5	merítsenek . Dávid ezen keresztül vette be a várát : " Azon a napon Dávid azt mondta : ' Akí	
doc#5	fölkapaszkodott a függőleges aknában (fönn a belső várból találta magát , s kinyitotta az ostromlók	
doc#5	másztak meg , amikor bevették a jebuzeusok várát . Az akna felső végében egy ferdén felfelé	
doc#5	karitatív intézete , neki adományozta Zólyom várát , s megbízta a Szepesség és a bányavárosok	
doc#5	Hunyadi Giskraval , kinek kezén maradtak várai és jövedelmei . V. László megerősítette	
doc#5	Mátyás csapatai mindenfelől támadták Giskra várait , s 1462 elején már csak Liptó , Zólyom	
Query vár, IGE.* 599,506 (482.6 per million)		
First	Previous	Page 73 of 23,981 Go Next Last
doc#7	ember keze nyújt , veszendő : Halhatatlant vár s keres a valódi , A nemes virtus : maga	
doc#7	uralkodónéja . " A nő közönségtől jótékony hatást vár a férfiúi nemre a hazai nyelv , az irodalom	
doc#7	székvárosát , mert ebben az időben Napóleontól várta a monarchia és hazája feudális viszonyainak	
doc#7	levegő árad ; olyasféle , amelyet már régen várt a kritika és a színház . A Rang és mód	
doc#7	egyik grófi származású , rá nagy örökség vár . Sem a nagypapa , sem a két lány , se	
doc#7	ellenezte az emigrációt is : tőle semmi jót nem várt , legfeljebb a Monarchia s vele együtt	
doc#7	ugyanígy vélekedik : " Kisült , amit nem vártunk , Mutatja a hatás : Rossz drámairó	
doc#7	még hátra maradt , leküzdeni - ... mi vár tehát ránk ? Azt a jó isten tudja , ember	
doc#7	foganatja lett . Nem volt más hátra , mint várni , és azt remélni , hogy az abszolutizmus	
doc#7	kudarca intő példa volt arra nézve , mi várna egy újabb magyar szabadságharcra . A kiegyezési	
doc#7	az emberiség minden kérdésének megoldását várták . A harmincas , negyvenes évek nemzedékeinek	
doc#7	mélyégeket , melyeknél halál és kárhozat vár ránk ! " Martinuzzi alakjánál hangsúlyosabbnak	

Irodalom

- ALEXIN, ZOLTÁN – GYIMÓTHY, TIBOR – HATVANI, CSABA – TIHANYI, LÁSZLÓ – CSIRIK, JÁNOS – BIBOK, KÁROLY – PRÓSZÉKY, GÁBOR (2003), Manually Annotated Hungarian Corpus. In: COPESTAKE, ANN – HAJIČ, JAN (eds), Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, East Stroudsburg. Vol. 2: 53–56.
- BARONI, MARCO – UEYAMA, MOTOKO (2006), Building General- and Special-purpose Corpora by Web Crawling. In: NOAMI, MARIKO (eds), Proceedings of the 13th NIJL International Symposium, Language Corpora: Their Compilation And Application. National Institute for Japanese Language, Tokyo. 31–40.
- BIEMANN, CHRIS – BILDHAUER, FELIX – EVERT, STEFAN – GOLDBAHN, DIRK – QUASTHOFF, UWE – SCHÄFER, ROLAND – SIMON, JOHANNES – SWIEZINSKI, LEO-

- NARD – ZESCH, TORSTEN. (2013), Scalable construction of high-quality web corpora. *Journal for Language Technology and Computational Linguistics* 28/2: 23–60.
- CSENDES, DÓRA – CSIRIK, JÁNOS – GYIMÓTHY, TIBOR (2004), The Szeged Corpus: A POS-tagged and Syntactically Annotated Hungarian Natural Language Corpus. In: SOJKA, PETR – KOPEČEK, IVAN – PALA, KAREL (eds), *Text, Speech and Dialogue*. (Lecture Notes in Computer Science, Vol. 3206) Springer, Berlin. 41–47.
- CSENDES, DÓRA – CSIRIK, JÁNOS – GYIMÓTHY, TIBOR – KOCSOR, ANDRÁS (2005), The Szeged Treebank. In: MATOUŠEK, VÁCLAV ET AL. (eds), *Proceedings of the 8th International Conference on Text, Speech and Dialogue*. Springer, Berlin. 123–31.
- ENDRÉDY ISTVÁN (2016) *Nyelvtechnológiai algoritmusok korpuszok automatikus építéséhez és pontosabb feldolgozásukhoz*. PhD disszertáció. Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar, Budapest.
- ENDRÉDY, ISTVÁN – INDIG, BALÁZS (2015), Huntag3: A General-Purpose, Modular Sequential Tagger – Chunking Phrases in English and Maximal NPs and NER for Hungarian. In: VETULANI, ZYGMUNT – MARIANI, JOSEPH (eds), *Proceedings of the 7th Language & Technology Conference, Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznań University, Poznań. 213–218.
- ENDRÉDY ISTVÁN – NOVÁK ATTILA (2012), Egy hatékonyabb webes sablonszűrő algoritmus – avagy miként lehet a cumisüveg potenciális veszélyforrás Obamára nézve. In: TANÁCS ATTILA – VINCZE VERONIKA (szerk.), IX. Magyar Számítógépes Nyelvészeti Konferencia. (MSZNY 2013). SZTE, Szeged. 297–301.
- ENDRÉDY, ISTVÁN – NOVÁK, ATTILA (2013), More Effective Boilerplate Removal – The GoldMiner Algorithm. *Polibits* 48: 79–83.
- ENDRÉDY ISTVÁN – NOVÁK ATTILA (2015), Szótövesítők összehasonlítása és alkalmazásai. *Alkalmazott Nyelvtudomány* 15/1–2: 7–27.
- HALÁCSY, PÉTER – KORNAI, ANDRÁS – NÉMETH, LÁSZLÓ – RUNG, ANDRÁS – SZAKADÁT, ISTVÁN – TRÓN, VIKTOR (2004), Creating Open Language Resources for Hungarian. In: LINO, MARIA TERESA – XAVIER, MARIA FRANCISCA – FERREIRA, FÁTIMA – COSTA, RUTE – SILVA, RAQUEL (eds), *Proceedings of 4th Conference on Language Resources and Evaluation*. ELRA, Paris. 203–210.
- KILGARRIFF, ADAM – BAISA, VIT – BUŠTA, JAN – JAKUBÍČEK, MILOŠ – KOVÁŘ, VOJTĚCH – MICHELFEIT, JAN – RYCHLY, PAVEL – SUCHOMEL, VÍT (2014), The Sketch Engine: Ten Years in Lexicography. *Lexicography*. *Journal of ASIALEX* 1/1: 7–36.
- LIGETI-NAGY NOÉMI (2015a), Szöveggörpuszok pontosabb annotációja gépi elemzéshez. In: BENŐ ATTILA – FAZEKAS EMESE – ZSEMLYEI BORBÁLA (szerk.), *Többsnyelvűség és kommunikáció Kelet-Közép-Európában*. BBTE, Kolozsvár. 421–429.
- LIGETI-NAGY, NOÉMI (2015b), Noun Phrase and What they Leave Behind—Rule-based NP-chunking in Hungarian Corpora. In: LIGETI-NAGY, NOÉMI (eds), *Computational Linguistic Methods in Applied Linguistics*. Jedlik Laboratories Reports, Vol. III. No. 5. Pázmány University ePress, Budapest. 35–57.
- MIHÁLTZ, MÁRTON – VÁRADI, TAMÁS – CSERTŐ, ISTVÁN – FÜLÖP, ÉVA – PÓLYA, TIBOR – KÖVÁGÓ, PÁL (2015), Beyond Sentiment: Social Psychological Analysis of Political Facebook Comments in Hungary. In: BALAHUR, ALEXANDRA ET AL. (eds),

- Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis. Association for Computational Linguistics, East Stroudsburg. 127–33.
- NEUNERDT, MELANIE – TREVISAN, BIANKA – CURY TEIXEIRA, TOMAS – MATHAR, RUDOLF – JAKOBS, EVA-MARIA (2011), Ontology-based Corpus Generation for Web Comment Analysis. In: DE BRA, PAUL (eds), Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia. ACM, New York. 335–336.
- ORAVECZ, CSABA – VÁRADI, TAMÁS – SASS, BÁLINT (2014), The Hungarian Gigaword Corpus. In: CALZOLARI, NICOLETTA ET AL. (eds), Proceedings of the Ninth International Conference on Language Resources and Evaluation. ELRA, Paris. 1719–1723.
- OROSZ, GYÖRGY – NOVÁK, ATTILA (2013), PurePos 2.0: A Hybrid Tool for Morphological Disambiguation. In: ANGELOVA, GALIA (ed.), Proceedings of the International Conference on Recent Advances in Natural Language Processing. INCOMA, Hissar. 539–545.
- POMIKÁLEK, JAN (2011), Removing Boilerplate and Duplicate Content from Web Corpora. PhD dissertation. Masaryk University, Faculty of Informatics, Brno.
- RECKSI, GÁBOR (2014), Hungarian Noun Phrase Extraction Using Rule-based and Hybrid Methods. *Acta Cybernetica* 21: 461–479.
- RECKSI GÁBOR – VARGA DÁNIEL (2012), Magyar főnévi csoportok azonosítása. *Általános Nyelvészeti Tanulmányok* 24: 81–95.
- PRÓSZÉKY GÁBOR – INDIG BALÁZS (2015), Magyar szövegek pszicholingvisztikai indítatású elemzése számítógéppel. *Alkalmazott Nyelvtudomány* 15/1–2: 29–44.
- RYCHLÝ, PAVEL (2007), Manatee/Bonito – A Modular Corpus Manager. In: SOJKA, PETR – HORÁK, ALEŠ (eds), Proceedings of Recent Advances in Slavonic Natural Language Processing (RASLAN 2007). Masaryk University, Brno. 65–70.
- TJONG, ERIK F. – SANG, KIM – VEENSTRA, JORN (1999), Representing text chunks. In: THOMPSON, HENRY S. – LASCARIDES, ALEX (eds), Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, East Stroudsburg. 173–79.
- VÁRADI, TAMÁS (2002), The Hungarian National Corpus. In: ZAMPOLLI, ANTONIO ET AL. (eds), Proceedings of the Third International Conference on Language Resources and Evaluation. ELRA, Paris. 385–389.

