Bevezetés a korpuszok és nyelvi adatbázisok világába

Vincze Veronika
tudományos főmunkatárs
MTA-SZTE Mesterséges Intelligencia Kutatócsoport
vinczev@inf.u-szeged.hu

Elméleti nyelvészetből és informatikatudományból doktoráltam a Szegedi Tudományegyetemen. Jelenleg számítógépes nyelvészként dolgozom az MTA-SZTE Mesterséges Intelligencia Kutatócsoportban, feladatom elsősorban a csoport projektjeinek nyelvészeti felügyelete és koordinálása. Érdeklődési körömbe elsődlegesen a korpuszépítés és a többszavas kifejezések számítógépes kezelése tartozik, de foglalkozom számítógépes morfológiával és szintaxissal, emellett információkinyeréssel is.

1. Bevezetés

A nyelvészeti kutatásokban jó ideje megkülönböztetik a kompetencia és performancia fogalmát (Chomsky 1957). Egy nyelv anyanyelvi beszélői kompetenciájuk segítségével képesek jól formált mondatokat alkotni az adott nyelven, így tudják eldönteni, hogy egy adott nyelvi megnyilatkozás megfelel-e a nyelv szabályainak vagy sem. A performancia ezzel szemben a nyelv gyakorlati megvalósulását jelenti: amit egy beszélő egy adott pillanatban kimond. Bizonyos esetekben a performancia nem követi a kompetenciát: ha például a beszélő fáradt vagy dekoncentrált, akkor elkövethet nyelvbotlásokat, megnyilatkozásában nem mindig követi a nyelv adott szabályait.

A nyelvészeti kutatások általában kétféle módszertannal dolgoznak: vannak adatorientált és elméletorientált módszerek. Az elméletorientált módszerek elsődlegesen a kompetenciára épülnek, azaz azt vizsgálják, az adott nyelvben mi lehetséges és mi nem, milyen szerkezetek lehetségesek és mik nem a nyelvi kompetenciának megfelelően. Vizsgálati módszereik igen gyakran épülnek introspekcióra, azaz a kutató a saját nyelvérzékére (intuíciójára) építve alkot lehetséges példamondatokat, melyeket aztán más anyanyelvi beszélőkkel véleményeztet, természetességüket, elfogadhatóságukat megítélendő.

Ezzel szemben az adatorientált módszerek a már létező nyelvi adatokból indulnak ki, ezeket elemzik, csoportosítják, ezeket próbálják meg szabályokkal leírni. A nyelvi adatokat a kutatók gyűjthetik adatközlőktől, például kérdőíves fel-

mérések vagy interjúk segítségével. Ezen felül a nyelvi adatok származhatnak adatbázisokból, szöveggyűjteményekből (azaz korpuszokból) is.

A korpusz ténylegesen előforduló írott vagy lejegyzett beszélt nyelvi adatok gyűjteménye. Általában speciális célokra hozzák létre őket, és a szövegek gyakran egy adott témakör köré csoportosulnak. A szövegeket valamilyen szempont szerint válogatják és rendezik. Nem feltétlenül egész szövegek vannak benne, és nem csak tárháza a szövegeknek, hanem sok esetben úgynevezett annotációt is tartalmaz: a szövegekben akár automatikus, akár kézi úton különféle nyelvi információk vannak jelölve, emellett a szövegek bibliográfiai adatai, szerkezeti egységei is eltárolódnak. A számítógépek kapacitásának megsokszorozódása révén a nagy méretű korpuszok összeállítása, tárolása és feldolgozása már megvalósítható, sőt kívánatos. A korpuszban található nyelvi adatok elemzése a korpusznyelvészet feladata.

E tanulmány célja, hogy az olvasót megismertesse néhány korpusszal és egyéb nyelvészeti adatbázissal, továbbá a korpusznyelvészet alapjaival. A legfontosabb alapfogalmak után ismertetjük a különféle korpusztípusokat, létrehozási módjukat, továbbá néhány példán keresztül megmutatjuk, milyen nyelvészeti jellegű információkat (annotációkat) tudunk a szövegekben kódolni. Arra is hozunk példát, hogy a nyers szövegállományból miként tudunk automatikusan annotált adatbázist előállítani. A korpuszok gyakorlati felhasználására is külön figyelmet fordítunk: bemutatjuk, hogy a korpuszokból származó adatokat hogyan lehetséges kigyűjteni, majd azokat nyelvészeti vagy más bölcsészettudományi kutatásra felhasználni.

2. Korpusztípusok

A korpuszokat számos szempont alapján csoportosíthatjuk: a szövegek nyelve szerint, modalitás szerint, a szövegek műfaja szerint stb. Modalitás szerint beszélhetünk írott nyelvi korpuszokról, melyek különféle szövegeket tartalmaznak, beszédkorpuszokról, melyek hanganyagokat és ezek szöveges átiratait foglalják magukban. Manapság pedig egyre nagyobb a multimodális korpuszok jelentősége is, melyek akár videófelvételeket is tartalmazhatnak, ezáltal hangi, képi és szöveges adatok is szerepelnek bennük.

Míg a korpuszok egy része egynyelvű dokumentumokból áll, addig számos korpusz két vagy több nyelven is tartalmaz(za ugyanazokat a) dokumentumokat. A párhuzamos korpuszokban ugyanannak a szövegállománynak többnyelvű

megfelelői vannak bekezdés, mondat és/vagy kifejezés szintjén megfeleltetve egymásnak: a világ egyik legnagyobb párhuzamos korpusza például a Biblia, melyet a világ számos nyelvére fordítottak már le.

A korpuszok – a szövegek tematikáját tekintve – lehetnek homogének, illetve heterogének, szintén az adott cél függvényében. Dönthetünk úgy, hogy minél nagyobb területet szeretnénk lefedni a nyelvi spektrumból, így több forrásból és témakörből választunk ki szövegeket. Ilyen például a Magyar Nemzeti Szövegtár (lásd lejjebb), amelynek készítői a sajtó, szépirodalom, tudományos, hivatalos és személyes stílusrétegéből válogattak szövegeket, odafigyelve arra is, hogy a határon túli nyelvváltozatok is képviselve legyenek a korpuszban. Ha azonban egy speciális alkalmazáshoz készítünk korpuszt, akkor igen gyakran behatárolt a témakör, például ha betegek dohányzási szokásait szeretnénk automatikusan kinyerni a kórlapokban rejlő információk alapján, magától értetődően orvosi jellegű dokumentumokat kell beépíteni a korpuszba. A szövegek kiválasztásakor arra is ügyelnünk kell, hogy az minél reprezentatívabb legyen az adott területre, azaz a szövegek rendelkezzenek a területre jellemző nyelvi és formai sajátosságokkal.

Beszélők vagy szerzők szerint is csoportosíthatjuk az adott szövegeket. Például egy korpusz tartalmazhat tájnyelvi szövegeket, ahol egy adott tájegységben élőktől származó nyelvi produktumokat gyűjtünk össze (például erdélyi magyar adatközlőktől gyűjtött szövegek). Fókuszálhatunk a gyermeki nyelvhasználatra a gyermeknyelvi korpuszok segítségével (lásd Babarczy 2019), illetve a nyelvtanulói korpuszokat használva felderíthetjük például a magyart mint idegen nyelvet tanulók számára nehezebb, problémásabb nyelvi jelenségeket (lásd Durst et al. 2013). Kitekintve más bölcsészettudományok felé, egy adott író vagy költő összes művei is tekinthetők egy írói korpusznak, lehetőséget adva mélyebb stilisztikai vagy egyéb irodalomtudományi elemzésekre.

Összeállíthatunk egy korpuszt egy adott nyelvi stílusréteg vagy regiszter szerint is, akár szakmai nyelvhasználatra való tekintettel is. Az utóbbira egy példa a Miskolc Jogi Korpusz vagy a SZEMEK orvosi szaknyelvi korpusz (Vincze 2018). Szempont lehet a szövegek kiválasztásában a szövegek keletkezései ideje, például a nyelvtörténeti, nyelvemlékeket tartalmazó korpuszok jöttek így létre (Simon 2019).

Természetesen léteznek olyan korpuszok is, melyek heterogén adatokat tartalmaznak, azaz több szövegtípusból, stílusrétegből és műfajból, valamint több szerzőtől származó szövegek is megjelennek az anyagban. Az ilyen általános célú

korpuszok esetében gyakran az adott nyelv vagy nyelvi réteg minél teljesebb reprezentációja a cél. Ezek a korpuszok sokszor nagyobb méretűek, jellemzően több millió szövegszót tartalmaznak, mint például a Magyar Nemzeti Szövegtár vagy a Szeged Korpusz (lásd lejjebb).

Az alábbiakban a teljesség igénye nélkül felsorolunk néhány ismertebb, az angol és a magyar nyelvre vonatkozó korpuszt.

A legnagyobb méretű, angol nyelvű szövegeket tartalmazó korpuszok az alábbiak: British National Corpus (BNC), Wall Street Journal (WSJ), Reuters. Ezek körülbelül 100 millió szövegszót tartalmaznak; a dokumentumok, bekezdések határai jelölve vannak bennük, egyéb (nyelvi) annotációt azonban nem foglalnak magukban. A Gigaword korpusz körülbelül 2 milliárd szóból áll, ez sem tartalmaz nyelvi annotációt – már méreténél fogva sem. A nyelvi annotációt tartalmazó angol nyelvű korpuszok közül a legismertebb a Penn TreeBank, mely 5 millió szövegszóból áll. A szavak szófaji kódja (POS-tag) meg van adva, és szintaktikai elemzés (konstituensfa) is található a korpusz mondataihoz.

A Magyar Nemzeti Szövegtár (Oravecz et al. 2014) a mai magyar írott köznyelv általános célú reprezentatív korpusza, amely a magyarországiak mellett a határon túli magyar nyelvváltozatokat is felöleli. Jelenleg több mint egymilliárd szövegszót tartalmaz. Az MNSZ lényegi tulajdonsága, hogy minden szó mellett feltünteti a szótövet, a szófajt és a szó morfológiai elemzését is. A szótő, szófaj és elemzés megállapítása és az elemzések egyértelműsítése automatikus gépi eszközökkel történik. A korpuszban való kereséshez külön online felület áll rendelkezésre (vö. Sass 2019).

A Szeged Korpusz és Treebank a legnagyobb, kézzel egyértelműsített magyar nyelvű adatbázis, melyben 1,2 millió szövegszó található hat különböző doménből (Csendes et al. 2005). A szövegek morfológiai és szintaktikai kézi elemzéssel rendelkeznek, valamint egyes részkorpuszokon további szemantikai annotációk (pl. tulajdonnevek) is elkészültek. A részletes kézi annotálásnak köszönhetően a treebank különböző verziói megbízható tanulási és tesztelési adatbázisként szolgálnak számítógépes tanulóalgoritmusok számára.

3. Annotáció

A legtöbb korpusz nem pusztán nyers szövegekből áll: általában be vannak jelölve a szöveg szerkezeti részei is, azaz szakaszokra, bekezdésekre, mondatokra, szövegszavakra (tokenekre) van bontva. Emellett többnyire annotációt is tartal-

maznak: az annotálási munkálatok során (nyelvész) szakértők – vagy automatikus annotáció esetében egy algoritmus – kézzel bejelölik a releváns információkat a szövegállományokban, például minden egyes szóhoz hozzárendelik a szófaját vagy a szövegben megjelölik a tulajdonneveket.

Az annotáció lehet dokumentumszintű (például egy e-mail spam-e, vagy sem), mondatszintű (például a mondat tényszerű információkat közöl-e, avagy bizonytalan, esetleg tagadott információt tartalmaz), illetve szószintű (például morfológiai elemzés). Egy korpuszban természetesen többféle annotáció is szerepelhet egyidejűleg, hiszen akár többszintű (morfológiai, szintaktikai és szemantikai) nyelvi elemzést is tartalmazhat egy adott korpusz. Mindemellett vannak annotáció nélküli korpuszok is: ezeket általában statisztikai célokra, például szógyakoriság megállapítására lehet hasznosítani (hányszor fordul elő egy adott szóalak egy kellően nagy korpuszban).

Az annotáció során (nyelvész) szakértők – előre meghatározott irányelvek alapján – kézzel bejelölik a szövegekben a releváns információkat, illetve ellenőrzik a gépi annotáció minőségét és kézzel javítják annak hibáit. Az annotálás módszertanát tekintve az annotáció lehet:

– egyszeres: egy szövegen egy annotátor megy végig;
– többszörös: egyazon szövegen több annotátor is teljes egészében végigmegy, egymástól függetlenül. Amennyiben eltérés mutatkozik a két (vagy több) annotáció között, egy újabb független annotátor dönt (egyértelműsít) a problémás esetekben.

A többszörös annotáció, noha időigényesebb és drágább, általában javítja az annotáció minőségét, hiszen több szakértő nézi át ugyanazt az anyagot. Előnyei közé tartozik még, hogy lehetővé teszi az egyetértési arány mérését is: az annotátorok által egyformán jelölt esetek százalékos arányát a gépi alkalmazások által elérhető felső határnak szokták tekinteni, így voltaképpen a feladat nehézségi fokának jelzésére is alkalmas ez a mérőszám. Az egyszeres annotáció előnyeként említhető, hogy olcsóbb és gyorsabb, mint a többszörös annotáció, azonban hátránya, hogy esetenként kevésbé pontos annotációt eredményez, és nem lehetséges vele egyetértési arányt mérni.

4. Korpuszépítés

Amennyiben nyelvészeti kutatásunkhoz korpuszból kívánunk adatokat gyűjteni, felmerül a kérdés, milyen korpuszt használjunk. Első kérdésként érdemes

megvizsgálnunk, hogy az adott kutatási témához illeszkedő korpusz elérhető-e számunkra. Ha rendelkezésre áll a céljainknak megfelelő korpusz, akkor elégséges lehet a meglevő korpuszból kigyűjteni a megfelelő adatokat. A korpuszban való keresési technikákról bővebben lásd Sass Bálint e kötetbeli tanulmányát (Sass 2019), a magyar nyelvű kereshető korpuszokról pedig a Nemzeti Korpuszportálon (http://corpus.nytud.hu/nkp) találunk bővebb információt.

Ha még korábban nem hoztak létre a céljainknak megfelelő korpuszt, akkor érdemes megfontolni a saját korpusz építését. Egy korpusz megtervezésekor és létrehozásakor számos szempontot kell mérlegelni. El kell döntenünk, hogy milyen célra kívánjuk használni a korpuszt – ennek ugyanis lényegi szerepe van a szövegek kiválasztásában, a korpusz méretének meghatározásában, az annotációs elvek kidolgozásában stb. Amennyiben a korpuszt tanító- vagy tesztadatbázisként szeretnénk hasznosítani algoritmusok fejlesztéséhez, fontos a megfelelő méret: elegendő nagynak kell ahhoz lennie, hogy kellő mennyiségű példát (és ellenpéldát) szolgáltasson az adott jelenségre. Az, hogy mi számít megfelelő méretnek, mindig az adott feladat függvénye: egy tulajdonnév-felismerő rendszer betanításához általában elegendő egy néhány százezer szövegszavas annotált korpusz (például Szeged NE korpusz, Szarvas et al. 2006), azonban egy szintaktikai elemző betanítása már milliós nagyságrendű szövegszóból álló annotált korpuszt igényel (például Szeged Treebank, Csendes et al. 2005).

A szövegek gyűjtéséhez el kell döntenünk a szövegek tematikáját (például jogi vagy irodalmi szövegeket szeretnénk vizsgálni). Döntést kell hozni a kutatni kívánt nyelvi regiszterekről is (például hivatalos nyelv, köznyelv, internetes nyelvhasználat…), valamint egyéb jellemzőkről is (például a szövegek keletkezési ideje vagy szerzője szerint is szűkíthetjük a kutatott szövegek halmazát). Nem elhanyagolható szempont a szövegek hozzáférhetősége sem, azaz egyrészt magunk hozzáférünk-e könnyen a korpuszba illeszteni kívánt szövegekhez, másrészt pedig hogy milyen módon tehetjük azokat hozzáférhetővé mások számára. Itt külön felhívnánk a figyelmet a szerzői jogokra – például irodalmi szövegek esetén –, illetve bizonyos szövegtípusok, különösen az orvosi és jogi dokumentumok megkövetelik a bennük szereplő érzékeny adatok anonimizálását.

A korpuszba bekerülő szövegek összegyűjtését azok gépi előfeldolgozása, illetve – amennyiben szükséges – digitalizálása követi. Az állományok automatikus megtisztítása, szakaszokra, bekezdésekre, mondatokra és tokenekre bontása után következhet az annotálási fázis (lásd részletesebben fent). A korpuszépítés utómunkálataiként megtörténik az annotált állományok összefésülése, a formai

hibák (automatikus és/vagy kézi) javítása, majd ezek után következhet a korpusz használatbavétele.

5. A korpuszok felhasználhatósága

A korpuszokat referencia-adatbázisként különböző alkalmazások tesztelésére szokás használni: a kézi annotációt etalonnak tekintve számszerűsíteni lehet, mennyire teljesít jól az adott rendszer (kiértékelés).

A tesztelés mellett a korpuszokat az algoritmusok betanítására is lehet használni. A tanítás során a szakértő példákat mutat az algoritmusnak az annotált korpuszból, amelyek alapján az algoritmus automatikusan állítja elő a szabályokat. Az algoritmus célja, hogy a tanult szabályok használatával a korábban nem látott példányokat is megtalálja / felismerje / azonosítsa. A feladattól függően számos példára lehet szükség a hatékony tanuláshoz.

6. Adatgyűjtés programozás nélkül: készítsünk szófelhőt!

Az alábbiakban bemutatjuk, hogyan tudunk könnyen és gyorsan látványos adatvizualizációt készíteni. Ehhez nincs szükség programozási tudásra, átlagos számítógép-felhasználói ismeretek segítségével is könnyen elboldogulunk.

Szerencsére már olyan elemző eszközök is rendelkezésre állnak, melyek programozói ismeretek nélkül is képesek támogatni a korpusznyelvészet iránt érdeklődőket. Az alábbiakban bemutatunk néhány olyan eszközt, melyek szövegek nyelvi elemzését, részletesebben: mondatra és szövegszavakra bontását, azok szófaji egyértelműsítését és morfológiai, valamint szintaktikai elemzését valósítják meg. E tanulmányban a magyarlanc és az UDPipe eszközöket mutatjuk be, de a kötetben Mittelholcz Iván tanulmánya részletesen is ismerteti az e-magyar eszközt, mely hasonló funkciókkal bír (Mittelholcz 2019).

A magyarlanc nevű nyelvi előfeldolgozó eszköz a Szegedi Tudományegyetem fejlesztése (Zsibrita et al. 2013). Egy magyar nyelvű szöveges állományból kiindulva (txt) képes a szöveg mondatokra és szavakra bontására, a szavak morfológiai elemzésére, majd szófaji egyértelműsítésére, továbbá kétféle szintaktikai elemzést is képes hozzárendelni a mondatokhoz, választhatóan függőségi (dependencia) nyelvtani elemzést vagy pedig összetevős elemzést. A magyarlanc elérhető a https://rgai.inf.u-szeged.hu/node/100 oldalon, az innen letölthető program segítségével txt formátumú szövegfájlok elemzése is lehetséges parancssorból. Ha pedig csak egy-egy mondat elemzésére van szükségünk, vagy pusztán

tesztelni szeretnénk az alkalmazást, erre a http://rgai.inf.u-szeged.hu/magyarlanc-service oldalon elérhető online demó nyújt lehetőséget.

A http://lindat.mff.cuni.cz/services/udpipe/ honlapon ingyenesen elérhető UDPipe nevű elemző a Universal Dependencies annotációs sémán alapul (Straka és Straková 2017), mely egy nemzetközileg egységes morfológiai és szintaktikai annotációs séma, jelenleg kb. 50 nyelvre – köztük magyarra – dolgozták ki. A magyarlanchoz hasonlóan képes a nyers szövegek mondatra és szavakra bontására és szófaji elemzésére, továbbá a mondatok függőségi elemzésére. Egy-egy mondat és szövegfájl elemzését egyaránt lehetséges elvégezni online a fenti honlapon.

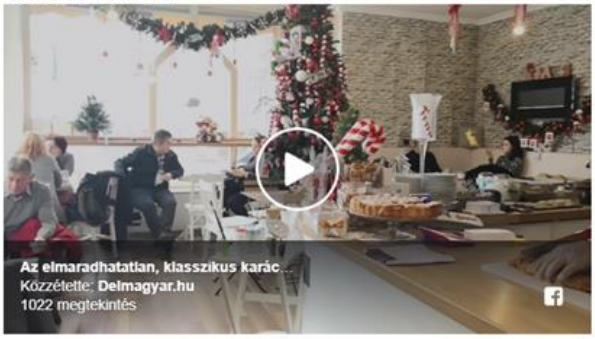
A két nyelvi elemző hasonló funkciókkal rendelkezik. Hogy a kettő közti választást elősegítsük, felsorolunk néhány további szempontot. Technikai oldalról talán könnyebben kezelhető az UDPipe, azonban kevéssé pontos elemzési eredményt ad a rendszer, mivel néhány ezer mondatnyi anyagon lett betanítva. Ezzel ellentétben a magyarlanc tanító anyaga kb. 70.000 mondatot tartalmaz, ami nagyságrendnyi különbséget jelent, és az elemzés pontosságára is kihatással van. Ugyanakkor a nemzetközi összevethetőség szemszögéből nézve az UD-sémára épülő elemzés többnyelvű vizsgálatok esetén hasznosabb lehet, mint a magyarlanc „magyarspecifikus” jegyekkel is bíró kimenete (megjegyezzük, hogy ez utóbbi szempont az UD és az e-magyar összevetésében is fennáll).

A továbbiakban megvizsgáljuk, hogyan tudunk programozói tudás nélkül is adatokat gyűjteni az elemzett fájlokból.

Első lépésben válasszunk ki egy nekünk szimpatikus szöveget! Ez lehet akár saját írásunk, akár az internetről gyűjtött szöveg, lényeg, hogy szöveges formátumban (txt) álljon rendelkezésünkre. Amennyiben egy internetes oldal tartalmát szeretnénk feldolgozni, illetve szöveges formátumban elmenteni, segítséget nyújthat a boilerpipe nevű eszköz. A https://boilerpipe-web.appspot.com oldalon elérhető eszköz megfelelő sorába illesszük be a letölteni kívánt oldal linkjét (legyen ez a példánkban a https://www.delmagyar.hu/szeged_hirek/kilometerekben_keszul_a_bejgli_-_az_elmaradhatatlan_klasszikus_karacsonyi_edesseget_kostoltuk/2583243 link), az Output Mode-ot állítsuk Plain textre, azaz sima szöveges állományra, majd nyomjunk az Extract gombra (2. ábra)! Ha a szöveges állomány még tartalmaz a weboldalról más fölösleges részleteket, kísérletezzünk azzal, hogy az Extractort is megváltoztatjuk, például LargestContentExtractorra vagy KeepEverythingExtractorra. Az 1. ábrán is látszik, hogy a weboldal eredetileg tartalmazott egy videót is, azonban a boilerpipe ezt nem exportálta szövegként.

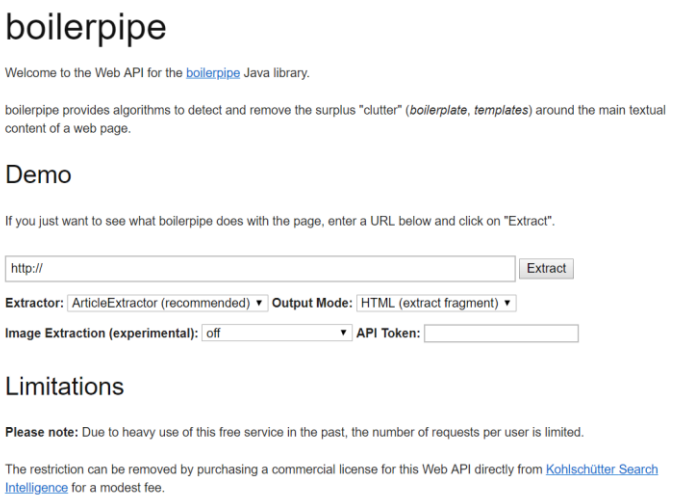
A bejgliket a Z. Nagy Cukrászdából, a Sugar & Candyből, az A Cappellából, a Reök Kézműves Cukrászda és Kávéházból, valamint a Lidlből és a Tescóból hoztuk.

A megjelenés alapján a négy cukrászdai termék átment a teszten. A Lidl bejglije méretével és kinézetével is kilógott a sorból. Az édesség az áruházláncnál 295 grammos és lapos. Viszont akciós és olcsó, 499 helyett mindössze 349 forint. A burritóhoz vagy kiflihez hasonló bejgliben ránézésre valóban sok a dió, de az ízén ez nem érződik.



A Tescóban árult édességet nem is merik bejglinek nevezni, hiszen az annak előállítására és minőségi követelményeire vonatkozó szabályokat a Magyar Élelmiszerkönyv tartalmazza, pontosan meg van határozva, milyen anyagokat lehet felhasználni a készítésükhöz, az elkészült termékeknek milyen kémiai, fizikai és érzékszervi tulajdonságokkal kell rendelkezniük. Ezeknek a tescós édesség nem felel meg, ezért omlós diós tekercs néven árulják, 400 grammot 499 forintért. Erre sem érdemes túl sok szót vesztegetni, de ebben legalább érződik a dió íze, viszont élvezhetetlenül száraz.

1. ábra: Egy online megjelent cikk



2. ábra: A boilerpipe online kezelőfelülete

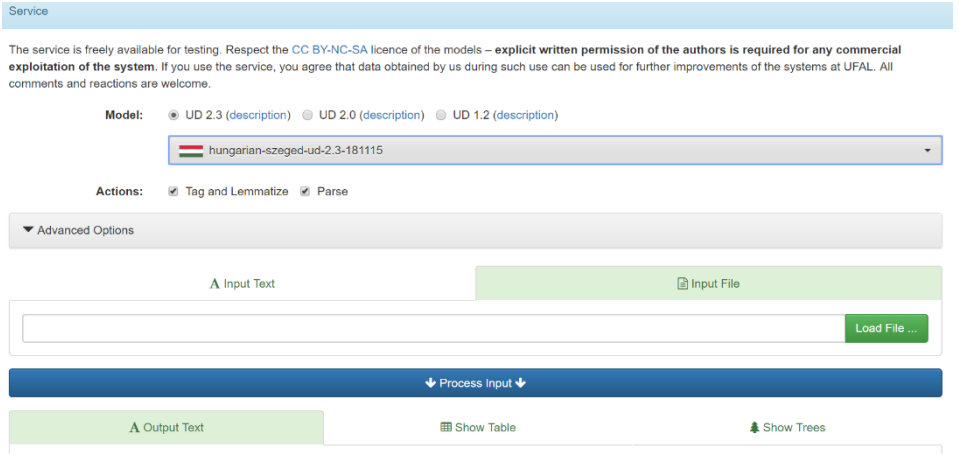
A kinyert tartalom egy részlete:

A bejgliket a Z. Nagy Cukrászdából, a Sugar & Candyből, az A Cappellából, a Reök Kézműves Cukrászda és Kávéházból, valamint a Lidlből és a Tescóból hoztuk.
A megjelenés alapján a négy cukrászdai termék átment a teszten. A Lidl bejglije méretével és kinézetével is kilógott a sorból. Az édesség az áruházláncnál 295 grammos és lapos. Viszont akciós és olcsó, 499 helyett mindössze 349 forint. A burritóhoz vagy kiflihez hasonló bejgliben ránézésre valóban sok a dió, de az ízén ez nem érződik.
A Tescóban árult édességet nem is merik bejglinek nevezni, hiszen az annak előállítására és minőségi követelményeire vonatkozó szabályokat a Magyar Élelmiszerkönyv tartalmazza, pontosan meg van határozva, milyen anyagokat lehet felhasználni a készítésükhöz, az elkészült termékeknek milyen kémiai, fizikai és érzékszervi tulajdonságokkal kell rendelkezniük. Ezeknek a tescós édesség nem felel meg, ezért omlós diós tekercs néven árulják, 400 grammot 499 forintért. Erre sem érdemes túl sok szót vesztegetni, de ebben legalább érződik a dió íze, viszont élvezhetetlenül száraz.

Amennyiben meg vagyunk elégedve a kinyert szöveges tartalommal, másoljuk ki a szöveget, és illesszük be egy szövegszerkesztőbe (Notepad vagy akár Microsoft Word), és szöveges állományként (txt) mentsük el!

Következő lépésként a mentett szöveget morfológiai és szintaktikai elemzésnek vetjük alá. Ehhez most a Universal Dependencies formalizmusra épülő UDPipe nevű eszközt használjuk fel, mely a http://lindat.mff.cuni.cz/services/udpipe/ oldalon érhető el (3. ábra). Először is minden más beállítást változatlanul hagyva válasszuk ki a magyar nyelvet, majd az Input file fülre kattintva a Load file gombbal válasszuk ki az előbbiekben elmentett txt fájlunkat! Ezután nyomjunk a Process input gombra! A Save output file gombra kattintva el tudjuk menteni az elemzett fájlt (4. ábra).

Keressük meg a fájlt a gépünkön, és szövegfájlként nyissuk meg például Notepadben! A teljes szöveg kimásolása után illesszük be az egészet egy üres Excel-munkafüzetbe (5. ábra)! Látjuk, hogy az eredeti szövegszavak a B oszlopban jelennek meg, továbbá ezek szótövesített alakjai a C oszlopot foglalják el, majd a D oszlop tartalmazza a szavak szófaját, az F az egyéb morfológiai jegyeket (például szám, személy, igeidő), végül a G és H oszlopok a függőségi elemzés



3. ábra: A UDPipe online kezelőfelülete

```
# sent_id = 2
# text = A megjelenés alapján a négy cukrászdai termék átment a teszten.
1   A   a   DET   _   Definite=Def|PronType=Art   2   det   _   _
2   megjelenés   megjelenés   NOUN   _   Case=Nom|Number=Sing|Number[psed]=None|Number[psor]=None|Person[psor]=None   3   nmod:att   _   _
3   alapján   alap   NOUN   _   Case=Sup|Number=Sing|Number[psed]=None|Number[psor]=Sing|Person[psor]=3   8   nmod:obl   _   _
4   a   a   DET   _   Definite=Def|PronType=Art   7   det   _   _
5   négy   négy   NUM   _   Case=Nom|Number=Sing|NumType=Card   6   amod:att   _   _
6   cukrászdai   cukrászdai   ADJ   _   Case=Nom|Degree=Pos|Number=Sing   7   amod:att   _   _
7   termék   termék   NOUN   _   Case=Nom|Number=Sing   8   nsubj   _   _
8   átment   átmegy   VERB   _   Definite=Ind|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act   0   root   _   _
9   a   a   DET   _   Definite=Def|PronType=Art   10   det   _   _
10   teszten   teszt   NOUN   _   Case=Sup|Number=Sing|Number[psed]=None|Number[psor]=None|Person[psor]=None   8   nmod:obl   _   SpaceAfter=No
11   .   .   PUNCT   _   _   8   punct   _   _
```

4. ábra: A morfológiailag és szintaktikailag elemzett szöveg egy részlete

# sent_id = 12									
# text = A megjelenés alapján a négy cukrászdai termék átment a teszten.									
1	A	a	DET	_	Definite=Def|PronType=Art	2	det	_	_
2	megjelenés	megjelenés	NOUN	_	Case=Nom|Number=Sing|Number[psed]=	3	nmod:att	_	_
3	alapján	alap	NOUN	_	Case=Sup|Number=Sing|Number[psed]=N	8	nmod:obl	_	_
4	a	a	DET	_	Definite=Def|PronType=Art	7	det	_	_
5	négy	négy	NUM	_	Case=Nom|Number=Sing|NumType=Card	6	amod:att	_	_
6	cukrászdai	cukrászdai	ADJ	_	Case=Nom|Degree=Pos|Number=Sing	7	amod:att	_	_
7	termék	termék	NOUN	_	Case=Nom|Number=Sing	8	nsubj	_	_
8	átment	átmegy	VERB	_	Definite=Ind|Mood=Ind|Number=Sing|Pe	0	root	_	_
9	a	a	DET	_	Definite=Def|PronType=Art	10	det	_	_
10	teszten	teszt	NOUN	_	Case=Sup|Number=Sing|Number[psed]=N	8	nmod:obl	_	SpaceAfter=No
11	.	.	PUNCT	_	_	8	punct	_	_

5. ábra: A morfológiailag és szintaktikailag elemzett szöveg egy részlete Excelben megjelenítve

részleteit takarják. Excel-szűrésekkel egyszerű statisztikai adatokat is tudunk gyűjteni, például: az adott szófajok aránya a szövegben, a leggyakoribb főnevek, amelyek alanyként szerepelnek a szövegben, tulajdonnevek a szövegben stb.

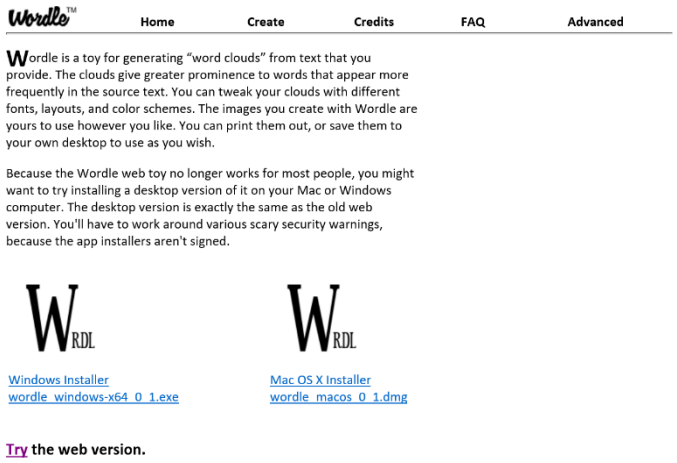
Tegyük fel, hogy a példában a szövegben megjelenő leggyakoribb főneveket szeretnénk egy szófelhő segítségével vizualizálni! Ehhez először is kapcsoljuk be

	A	B	C	D	E	F	G	H	I	J	K
1	# newd										
6	2	hírei	hír	NOUN	_	Case=Nom|Number=Plur|Number[psed]=	5	nsubj	_	_	
8	4	Kilométerekben	Kilométerek	NOUN	_	Case=Ine|Number=Plur	5	nmod:obl	_	_	
18	14	édességet	édesség	NOUN	_	Case=Acc|Number=Sing	15	obj	_	_	
20	16	Kilométerekben	Kilométerek	NOUN	_	Case=Ine|Number=Plur	17	nmod:obl	_	_	
30	26	édességet	édesség	NOUN	_	Case=Acc|Number=Sing|Number[psed]=N	27	obj	_	_	
50	1	Diósból	Diós	NOUN	_	Case=Ela|Number=Sing|Number[psed]=N	2	nmod:obl	_	_	
52	3	hatot	hat	NOUN	_	Case=Acc|Number=Sing|Number[psed]=N	4	obj	_	_	
51	3	mindenkit	mindenk	NOUN	_	Case=Acc|Number=Sing|Number[psed]=N	4	obj	_	_	
56	8	ízével	íz	NOUN	_	Case=Ins|Number=Sing|Number[psed]=N	4	nmod:obl	_	SpaceAfter=No	
70	12	árával	ár	NOUN	_	Case=Ins|Number=Sing|Number[psor]=Sir	8	conj	_	_	
78	3	halászlé	halászlé	NOUN	_	Case=Nom|Number=Sing|Number[psed]=	11	obl	_	_	
89	14	asztalról	asztal	NOUN	_	Case=Del|Number=Sing|Number[psed]=N	15	nmod:obl	_	_	
96	2	cukrászdák	cukrászda	NOUN	_	Case=Nom|Number=Plur|Number[psed]=	16	obl	_	SpaceAfter=No	
98	4	pékségek	pékség	NOUN	_	Case=Nom|Number=Plur|Number[psed]=	2	conj	_	_	
01	7	bevásárlóközpontok	bevásárlóközpont	NOUN	_	Case=Nom|Number=Plur|Number[psed]=	13	nmod:att	_	SpaceAfter=No	
03	9	áruházak	áruház	NOUN	_	Case=Nom|Number=Plur|Number[psed]=	7	conj	_	_	
06	12	kisboltok	kisbolt	NOUN	_	Case=Nom|Number=Plur|Number[psed]=	7	conj	_	_	
07	13	polcai	polc	NOUN	_	Case=Nom|Number=Plur|Number[psed]=	16	nsubj	_	_	
09	15	bejglitől	bejgli	NOUN	_	Case=Abl|Number=Sing	16	nmod:obl	_	_	

6. ábra: Az elemzett szövegből leszűrt főnevek

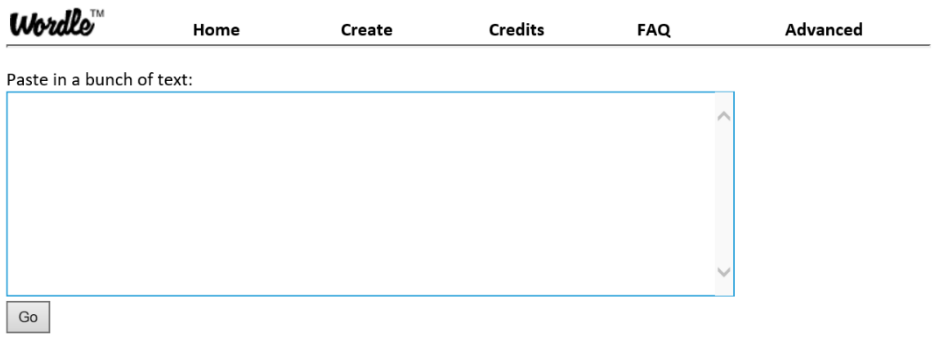
az Excel szűrő funkcióját, és a D oszlopból gyűjtsük ki a főneveket (NOUN) (6. ábra)! A szűrt sorokban jelöljük ki a C oszlopot (feltételezve, hogy a szótövesítés utáni alakok gyakorisága érdekel minket, tehát a *bejglit*, *bejglivel* stb. alakokat egyként (*bejgli*) szeretnénk kezelni). Az így kapott szólistát fogjuk vizualizálni a Wordle program segítségével.

Ehhez nyissuk meg a http://www.wordle.net oldalt, itt kattintsunk a Try the web version opcióra (7. ábra)! (Ha nem működik, akkor érdemes letölteni a programnak az operációs rendszerünknek megfelelő asztali verzióját, majd feltelepíteni azt, az utasításokat követve.) Amennyiben működik a webes változat, illesszük be az előzőekben az Excelből leszűrt főnévlistát, majd kattintsunk a Go gombra (8. ábra)! Eredményül egy szófelhőt kell kapnunk, melyen a betűméret



7. ábra: A Wordle online kezelőfelülete

jelöli az előfordulási gyakoriságot, tehát minél nagyobb betűkkel jelenik meg egy szó, annál gyakrabban fordult elő szövegünkben (9. ábra). A mi példánkban a *bejgli* szó tűnik a leggyakoribbnak, de az *édesség*, *forint*, *íz* és *diós* szavak is sokszor fordultak elő.



8. ábra: Szöveg beillesztése a Wordle-be

9. ábra: Az újságcikk leggyakoribb főnevei, a Wordle segítségével megjelenítve

A Font, Layout és Color menüpontokban igény szerint szabadon változtathatjuk az ábra színeit, betűtípusát, a Language menüpontban pedig be tudjuk állítani, hogy a leggyakoribb nyelvtani szavakat (az úgynevezett stopszavakat, mint például *és*, *vagy*, *a*, *ez*, *az*, *van*…) figyelembe vegye-e a program. Lehetőségünk van a szófelhő elmentésére és kinyomtatására is.

7. Összegzés

E tanulmányban röviden bemutattuk a korpuszok jelentőségét a nyelvészeti kutatásban, valamint ismertettünk néhány önálló, programozási tudást nem igénylő módszert, melyek segítségével a korpuszokból adatokat tudunk gyűjteni, illetve azokat elemezni. A tanulmánynak nem lehetett célja a teljes részletességre törekvés sem a módszerek, sem a korpuszok ismertetésekor, azonban az érdeklődő olvasó számára az alábbiakban szeretnénk néhány további lehetséges irányt felvázolni.

A korpusznyelvészetről részletes áttekintést nyújt Szirmai Monika könyve (Szirmai 2006). Az elemzett korpuszokban való kereséshez, különös tekintettel a Magyar Nemzeti Szövegtárra, Sass Bálint e kötetbeli tanulmánya mutat be különböző módszereket (Sass 2019), illetve a Nemzeti Korpuszportálon összegyűjtött korpuszokban is lehetséges adatokat keresni. A történeti korpuszokról Simon Eszter, a gyermeknyelvi korpuszokról Babarczy Anna tanulmányában olvashatunk részletesen (Simon 2019, Babarczy 2019). Végül egy további alkalmazást is szeretnénk az olvasó figyelmébe ajánlani: a TANIT online szolgáltatás a magyarlanc elemzéseire építve képes a szövegre jellemző alapvető statisztikai adatokat automatikusan összegyűjteni (lásd Péter 2019). Akit pedig mélyebben érdekel a programozás, Hammond *Java for Linguists* című könyvéből elsajátíthatja a nyelvészeti kutatáshoz szükséges programozás alapjait (Hammond 2002).

Irodalom

Babarczy A. 2019. Gyermeknyelvi korpuszok és erőforrások. In: Sulyok H., Juhász V., Erdei T. (szerk.). *Beszéd- és nyelvelemző szoftverek a versenyképességért és az esélyegyenlőségért. HunCLARIN korpuszok és nyelvtechnológiai eszközök a bölcsészet- és társadalomtudományokban*. Szeged: SZTE JGYPK Magyar és Alkalmazott Nyelvészeti Tanszék.

Chomsky, N. 1957. *Syntactic Structures*. The Hague: Mouton and Co.

Csendes D., Csirik J., Gyimóthy T., Kocsor A. 2005. The Szeged Treebank. In: Matoušek, V. et al. (szerk.). *Proceedings of the 8th International Conference on Text, Speech and Dialogue (TSD 2005)*. Berlin, Heidelberg: Springer-Verlag. 123–131.

Durst P., Szabó M. K., Vincze V., Zsibrita J. 2013. A HunLearner magyar tanulói korpusz fejlesztése és várható hozadékai. *THL2: A magyar nyelv és kultúra tanításának szakfolyóirata* 9/1–2. 28–41.

Hammond, M. 2002. *Programming for linguists: Java™ technology for language researchers*. Oxford: Blackwell.

Mittelholcz I. 2019. Bevezetés az `e-magyar` programcsomag használatába. In: Sulyok H., Juhász V., Erdei T. (szerk.). *Beszéd- és nyelvelemző szoftverek a versenyképességért és az esélyegyenlőségért*. *HunCLARIN korpuszok és nyelvtechnológiai eszközök a bölcsészet- és társadalomtudományokban*. Szeged: SZTE JGYPK Magyar és Alkalmazott Nyelvészeti Tanszék.

Oravecz Cs., Váradi T., Sass B. 2014. The Hungarian Gigaword Corpus. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC2014)*. Reykjavík: European Language Resources Association.

Péter R. 2019. A big data kihívás a bölcsészettudományokban: néhány digitális bölcsészeti kutatási eszköz bemutatása. In: Sulyok H., Juhász V., Erdei T. (szerk.). *Beszéd- és nyelvelemző szoftverek a versenyképességért és az esélyegyenlőségért*. *HunCLARIN korpuszok és nyelvtechnológiai eszközök a bölcsészet- és társadalomtudományokban*. Szeged: SZTE JGYPK Magyar és Alkalmazott Nyelvészeti Tanszék.

Sass B. 2019. Keresés korpuszban 2: így kerestek ti. In: Sulyok H., Juhász V., Erdei T. (szerk.). *Beszéd- és nyelvelemző szoftverek a versenyképességért és az esélyegyenlőségért*. *HunCLARIN korpuszok és nyelvtechnológiai eszközök a bölcsészet- és társadalomtudományokban*. Szeged: SZTE JGYPK Magyar és Alkalmazott Nyelvészeti Tanszék.

Simon E. 2019. Magyar nyelvű történeti korpuszok. In: Sulyok H., Juhász V., Erdei T. (szerk.). *Beszéd- és nyelvelemző szoftverek a versenyképességért és az esélyegyenlőségért*. *HunCLARIN korpuszok és nyelvtechnológiai eszközök a bölcsészet- és társadalomtudományokban*. Szeged: SZTE JGYPK Magyar és Alkalmazott Nyelvészeti Tanszék.

Straka, M., Straková, J. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver: Association for Computational Linguistics. 88–99.

Szarvas Gy., Farkas R., Kocsor A. 2006. A Multilingual Named Entity Recognition System Using Boosting and C4.5 Decision Tree Learning Algorithms. In: *Discovery Science 2006*. Berlin, Heidelberg: Springer-Verlag 267–278.

Szirmai M. 2006. *Bevezetés a korpusznyelvészetbe*. *A korpusznyelvészet alkalmazása az anyanyelv és az idegen nyelv tanulásában és tanításában*. Budapest: Tinta Kiadó.

Vincze V. 2018. A Miskolc Jogi Korpusz nyelvi jellemzői. In: Szabó M., Vinnai E. (szerk.). *A törvény szavai: Az OTKA-112172 kutatási zárókonferencia anyaga*. Miskolc: Bíbor Kiadó. 9–36.

Zsibrita J., Vincze V., Farkas R. 2013. magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: *Proceedings of RANLP 2013*. Hissar: Association for Computational Linguistics. 763–771.