Tulajdonnevek felismerése az Európai Média Monitor magyar moduljának fejlesztésében

1. Bevezetés. A Közös Kutatóközpont – Európa (European Joint Research Centre) által fejlesztett európai médiafigyelő (http://emm.newsexplorer.eu) világszerte több ezer hírportálról, 75 nyelvből automatikusan gyűjti, és különféle kategóriákba sorolja a híreket, a nap 24 órájában, tízpercenként frissítve, nyelvtechnológiai eszköztár használatával. A főmenüben láthatók a legtöbb portálon megjelent hírcsoportok, az ún. klaszterek. Minden, a vizsgált hírportálokon feltüntetett hírt automatikusan feldolgoznak, tartalmukra való tekintet nélkül.

Ha ugyanaz a hír több nyelven is megjelenik, szintén nyelvtechnológiai módszerekkel (pl. gépi fordítással) igyekszenek felismerni a hír azonosságát. Ennek köszönhető a rendszer egyik legizgalmasabb szolgáltatása: lehetőség ugyanazon hír különböző nyelveken közölt változatainak az összehasonlítására. Ez tekinthető egyfajta korlátozott párhuzamos korpusz alapanyagának is. Azért korlátozott, mert maguk a hírek hosszabb távon nem elérhetők, nincsenek archiválva. Csupán az aktuálisan elérhető hírekből szemezgethetünk, ha alaposabban meg szeretnénk nézni, milyen hasonlóságok és különbségek vannak ugyannak a hírnek akár azonos, akár különböző nyelvű változatai között.

A híreket a bennük szereplő földrajzi neveket használva a világtérképen is elhelyezi a rendszer. A gyakori személynevekről és fontosabb intézménynevekről is találunk összesítést, áttekinthetjük a személyek kapcsolatrendszerét, megnézhetjük, különféle nyelveken milyen írásmóddal szerepelt az adott név, és az egyes személyekről szóló legfrissebb híreket is megjeleníthetjük. A személyek által és a róluk mondott idézeteket is gyűjti a rendszer. Az EMM NewsExplorer oldalról, amely az előző napi fontosabb híreket foglalja össze, könnyedén átjuthatunk az EMM NewsBrief oldalra, amely az éppen aktuális aznapi híreket mutatja. Ennek az oldalnak az elején, grafikonon is megjelenítik, hogyan terjed és cseng le egy-egy hír az általuk vizsgált portálokon.

A kiemelt témakörökhöz tartozó híreket tematikus csoportokba sorolva is áttekinthetjük: például bűnözés, terrorizmus, emberek által okozott katasztrófák, konfliktusok. Ha témakörökre keresünk, nem csupán az aznapi, hanem a témakörhöz tartozó korábbi hírekbe is belenézhetünk. Beállíthatjuk, mely országról szóló híreket szeretnénk látni, és azt is, hogy milyen nyelvű hírekre vagyunk kíváncsiak. A teljes rendszert, amelynek első, még jóval kevesebb nyelvre alkalmazott változatát 2002-ben kezdték üzemeltetni, STEINBERGER ismerteti szerzőtársaival (2009).

A magyar nyelvű változat kidolgozásához elsősorban magam készítettem el a szükséges forrásanyagokat, és ellenőriztem a tesztek eredményeit. A különféle listák összeállításához PINTÉR TIBOR és SIMON ESZTER szolgáltatott alapanyagot, a munkálat irányítója a Nyelvtudományi Intézet részéről VÁRADI TAMÁS. Annak érdekében, hogy a szolgáltatások a magyar nyelvre is működjenek, mindenekelőtt javaslatot tettem a figyelembe veendő magyar nyelvű hírportálokra. Így a legolvasottabb hazai hírportálokon túlmenően regionális és határon túli portálokat is feldolgoz a rendszer. Mivel emberi beavatkozás

DOI: https://doi.org/10.29178/NevtErt.2015.17 Névtani Értesítő 37. 2015: 239–244.

nélkül, különféle politikai beállítottságú portálokról gyűlnek a hírek, nyugodt lelkiismerettel állíthatjuk, hogy a legkevésbé elfogult hírforrás, amelynek külön előnye, hogy hirdetések nem szerepelnek benne. A magyar nyelv beillesztése a rendszerbe gazdag morfológiája miatt jelentett fokozott kihívást. A problémákról és megoldásukról részleteiben több tanulmány is beszámolt (STEINBERGER et al. 2013, PAJZS 2014, PAJZS et al. 2014). A magyar modul 2013 óta működik. Általánosságban a tulajdonnevek automatikus felismerésének problémáival és megoldásuk lehetőségeivel részletesen SIMON ESZTER foglalkozott PhD-értekezésében (2013).

2. A személynevek felismerése. A NewsExplorer számára fejlesztett többnyelvű személynév-felismerő rendszer az alábbi segédanyagokat használja, amelyeket nyelvenként egyesével készítenek el (a magyar változat saját munkám): titulusok (*őfelsége, úr, asszony* stb.), kiemelt pozíciók (*miniszterelnök, kormányzó* stb.), foglalkozások (*festőművész, sebész, biológus* stb.) népnevek (*finn, francia, magyar, baszk* stb.) listája. Ezek segítenek az olyan szerkezet felismerésében, mint *Manuel Valls francia miniszterelnök; dr. Jari Louhelainen finn biológus; Wilhem von Humboldt német filozófus; Kelemen Attila erdélyi újságíró*. A rendszer által korábban nem ismert, legalább két elemből álló egységeket (egymást követő, nagybetűvel kezdődő szavakat) akkor értelmezik személynévként, ha az utolsó elem egy magyar keresztnév, akár toldalékolt formában, és a fenti listák valamelyikéből legalább egy elem szerepel közvetlenül a névként értelmezhető elem mellett. Így az alábbi példamondatrészletből Vámos Tibor nevét nem ismeri fel személynévként, a többit igen: *vendége lesz mások mellett Ungvári Tamás professzor, Vámos Tibor és Csányi Vilmos akadémikus, Bodor Pál író-újságíró*. Ha azonban egy másik hírben (vagy akár ugyanennek a hírnek egy másik portálon eltérően megfogalmazott változatában) feltűnik a *Vámos Tibor akadémikus* szerkezet, akkor az ő neve is felkerül az ismert személyek listájára, és egy újabb, hasonló mondatszerkezetben már az *akadémikus* szó szerepeltetése nélkül is felismeri a rendszer.

A magyar nyelvű modul számára a személynevek felismertetéséhez a fenti listák elkészítésén túlmenően a következőkre kellett figyelemmel lennünk:

a) A nemzetközi személynevek előfordulhatnak a magyar hírekben keresztnév vagy egyénnév + vezetéknév, illetve vezetéknév + keresztnév vagy egyénnév formában is, sőt gyakran csak a vezetéknév szerepel. Ráadásul a változatosságra törekvés jegyében egy híren belül sokszor hol a teljes nevet, hol csak a vezetéknevet, esetenként csak a keresztnevet vagy egyénnevet használják. Nők esetében még többféleképpen hivatkoznak akár egy híren belül ugyanarra a személyre: a hivatalos *Hillary Clinton* formán kívül előfordulhat a *Mrs.* vagy *Ms. Clinton,* sőt a *Clintonné* változat is, de még a puszta *Hillary* sem kizárt. A jelenleg megvalósított változat magyar szövegekben a fenti példákból csak a *Hillary Clinton* változatot ismeri fel, mivel a tesztelések közben úgy találtuk, túl sok félreértést okozhat a különböző változatok kezelésének a kísérlete. A változatosság több problémát is okozhat az EMM rendszer számára, hiszen egyebek között arra törekszenek, hogy nyilvántartsák, ugyanaz a személy hányszor és hol szerepelt a hírekben, továbbá milyen idézetek köthetők a nevéhez.

b) A nemzetközi nevek egy része a magyar toldalékok előtt megváltozik *(Obamát, Ronaldóról)*.

c) A nemzetközi neveket esetenként átírjuk a magyar helyesírásnak megfelelően, de ugyanaz a név másik hírben esetleg szerepelhet más formában is (pl. *Putyin ~ Putin, Dobrovolszkij ~ Dobrovoslki*).

d) A foglalkozásnevek, titulusok stb. listájában szereplő szótövek is változhatnak toldalékolt formájukban *(szakáccsal, őfelségét).*

A felsorolt nehézségek ellenére a rendszer jól azonosítja az egyes személynevek különféle nyelveken írt különböző változatait (pl. *Angela Merkel, Frau Merkel, Angela Merckel, Chancellor Merkel, Angela Dorothea Merkel, Niemiec Angela Merkel, Angela Merkelová* stb.) A magyar hírekben a toldalékolt nevek helyes felismerését többféleképpen teszteltük. A nehézséget főként az okozza, hogy az EMM rendszerbe lehetőleg nem kívánnak beilleszteni nyelvfüggő elemeket, így a létező magyar morfológiai elemzők egyikét sem szerették volna használni. Megadtuk a leggyakoribb névszói toldalékok listáját, ezután kipróbáltuk, mi történik, ha ezeket automatikusan levágják azokról az elemekről, amelyeket személynévnek véltek annak alapján, hogy nagybetűvel kezdődött, és mellette volt a felsorolt elemek valamelyike (titulus, foglalkozás, beosztás). Ezzel a módszerrel azonban a *Gordon* keresztnévből levágták az *-on* feltételezett toldalékot, a *Csaba* névből a feltételezett *-ba* toldalékot, vagy akár *Borisz Paszternak* nevéből a *-nak* feltételezett toldalékot. Ezután teszteltük, mi történik, ha nem vágják le a toldaléknak tűnő elemeket. Az ellenőrzésnél úgy láttuk, lényegesen megnyugtatóbb az eredmény, mivel az igazán fontos, gyakran előforduló személynevek általában toldalék nélkül szerepelnek a szövegben: egyrészt azért, mert azokat a személyeket jelölik, akik mondanak, csinálnak valami említésre méltót, ezért többnyire alanyesetben van a nevük, másrészt azért, mert ha a név mellett szerepel a foglalkozásukat vagy pozíciójukat jelölő szó, az esetleges toldalék ez utóbbira kerül. Így az esetek több mint 98%-ában toldalék nélkül szerepelnek a gyakori nevek a hírekben. A tesztelés során azt is megállapítottuk, mely toldalékok fordulnak elő leggyakrabban a személynevek után: a *-t,* a *-nAk* és a *–vAl,* a toldalékolt alakok 68%-ában. A rendszer végleges változata már csak a gyakori toldalékokat vágja le az ismert nevekről.

A tulajdonnév-felismerés pontosításához átadtunk egy ún. stopword listát is: ezek azok a kötőszavak, határozók, amelyek gyakran fordulnak elő mondat elején, így nagybetűvel kezdődnek, és a program számára személynév elemének tűnhetnek (*Bár, Ezért, Ha, Mi, Mit, Most, Mindeközben* stb.). Ezeket így már nem tekinti név elemének a program. Természetesen ez is okozhat félreelemzéseket: például *Mit Romney* neve megcsonkulhat. Az ilyen típusú félreelemezés részben elkerülhető, ha előbb azt nézik meg, a feltételezett személynév előfordul-e a már ismert személynevek között, és csak akkor használják a stopword listát, ha a személynévnek tűnő elem nem fordul elő az ismert nevek listájában. Mivel az aktuális hírekből gyűlnek a személynévlisták, történelmi személyek csak akkor kerülnek fel a listára, ha az elmúlt időszak híreiben szerepeltek. Ilyen módon azonban elvileg kitalált személyek nevei is felkerülhetnek a listára.

3. Földrajzi nevek felismerése. Az EMM földrajzi névi adatbázisát több adatbázisból állították össze. Ezek egyike a KNAB többnyelvű adatbázisa (http: //www.eki.ee/knab/p_mm_en.htm), amely az egyes földrajzi nevek különböző nyelvű változatait tartalmazza (*Venice ~ Venise ~ Venedig, Constantinapol ~ Istanbul*). Az EMM adatbázis összeállításának részleteiről POLIQUEN et al. (2006) számol be. Az adatbázis első mezője a megnevezett hely, helység, járás, megye, ország, víz, hegy, tájegység földrajzi koordinátáit

tartalmazza, a második mező azt az információt, hogy milyen jellegű földrajz névről van szó (ország, nagyváros, kisebb település stb.), a következő mező tartalmazza a hely angol nevét, majd ezt követik a különböző nyelvű és írásmódú változatok. Egyebek közt határon túli nevek magyar változatait is tartalmazza az adatbázis *(Brezno, Breznó),* a teljesség igénye nélkül. Ahol hiányt észleltem, kiegészítettem a sort a nemzetközi nevek magyar változataival *(Wien ~ Bécs, Being ~ Peking, Venice ~ Velence, Brussels ~ Brüsszel)* és a toldalékok előtt előforduló tőváltozatokkal *(Velencé-, Ausztriá-)*. Egyes neveknél számos változatot tüntettem fel: *Malta Island* | *Malta Island%* | *Málta szigete* | *Málta szigeté%* | *Málta-szigete* | *Málta-szigeté%* | *Málta* | *Máltá%* . A % jel itt azt jelöli, hogy a nevet követheti valamilyen toldalék, a | jel pedig az egyes változatokat különíti el. Így a toldalékolt földrajzi nevek is felismerhetők. Tapasztalatunk szerint a feldolgozott hírekben a földrajzi nevek 40%-a toldalékolt formában fordult elő. Cseppet sem meglepő módon az elsöprően gyakran – a toldalékolt alakok több mint 70%-ában – használt toldalék a *-bAn* és az *-On* volt.

A program a helyesen felismert földrajzi nevek alapján tudja elhelyezni a világtérképen az egyes híreket. Előfordulnak azért az automatikus eljárásból adódó mulatságos félreértések is: például a *Gabonasiló* szót tartalmazó hírt Gabon államhoz rendelte a program, a *Drámai* szót tartalmazót pedig Indiába, ahol *Dráma* nevű helységet talált. A hasonló, gyakrabban előforduló hibákra lehetőséget adó magyar szavakat egy ún. geo-stopword listába gyűjtöttük; ezt használva a program már lényegesen kevesebb, mindössze 3%-nyi hibával helyezte el a híreket a világtérképen.

4. Intézménynév, egyéb tulajdonnév. A fontosabb nyelvekre már kidolgozták az intézmények, szervezetek stb. neveinek az automatikus felismerését is, így az angol változatban szerepel például a *United Nations, NATO, Pentagon, Amnesty International,* és megtekinthetjük az ezekkel az intézményekkel leggyakrabban együtt emlegetett személyek neveit is. A magyar cikkekben a rendszer a már ismert nemzetközi intézmények neveit azonosítja természetesen a legkönnyebben, de már felismeri a *Magyar Labdarúgó Szövetség* vagy a *Magyar Nemzet* napilap nevét is, és természetesen – az angol intézménynevekhez hasonlóan – meg tudja mutatni a hozzájuk leggyakrabban kapcsolódó, felismert személyneveket is. A szélesebb körű magyar intézménynév-felismeréshez szolgáltattunk ugyan segédanyagokat (oktatási intézmények, színházak, egyházak, könyvtárak, civil szervezetek stb. aktuális listái), ezek beillesztését a rendszerbe azonban későbbre halasztották, mivel már a jelenlegi soknyelvű változat is nagy kihívást jelent a rendszer üzembiztos működése szempontjából.

5. A Média Monitor egyéves működése utáni tapasztalatok összegzése. A 2012. december – 2013. augusztus között eltelt időszakban több mint 100 000 személynevet ismert fel a rendszer; ezeknek kevesebb mint 10%-a volt toldalékolt formában. A leggyakrabban előforduló személyneveknek mindössze 1,8%-a volt toldalékolt, így az EMM előző napi híreket összesítő listáján, ahol a leggyakrabban említett neveket tüntetik fel, csak elvétve fordul elő toldalékolt név.

A különböző tulajdonnévlisták kiértékelésének melléktermékeként toldalékstatisztika is készült. Ezek alapján megfontolásra érdemesnek tűnt, hogy csupán a leggyakoribb toldalékok felismerését célozzuk meg. Míg a személyneveknél a *-t* tárgyrag különböző alakjai, valamint a *-nAk* és a *-vAl* ragok fordultak elő leggyakrabban, addig a földrajzi

neveknél főként a *-bAn* és az *-On* toldalékok alakváltozatai szerepeltek. A személynevekhez rendelt felismert idézetek száma átlagosan naponta 20. Összességében elmondható, hogy az EMM magyar modulja működőképes.

Hivatkozott irodalom

PAJZS JÚLIA 2014. Az Európai Médiafigyelő (EMM) magyar változata. In: TANÁCS ATTILA – VARGA VIKTOR – VINCZE VERONIKA szerk., *MSZNY 2014. X. Magyar Számítógépes Nyelvészeti Konferencia (Szeged, 2014. január 16–17).* Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged. 259–268.

PAJZS, JÚLIA – STEINBERGER, RALF – MAUD, EHRMANN – MOHAMED, EBRAHIM – DELLA ROCCA, LEONIDA – BUCCI, STEFANO – SIMON, ESZTER – VÁRADI, TAMÁS 2014. Media Monitoring and Information Extraction for the Highly Inflected Agglutinative Language Hungarian. In: CALZOLARI, NIKOLETTA – CHOUKRI, KHALID – DECKLERCK, THIERRY – LOFTSSON, HRAFN – MAEGAARD, BENTE – MARIANI, JOSEPH – MORENO, ASUNCTION – ODIJK, JAN – PIPERIDIS, STELIOS eds., *LREC 2014. Proceedings of the Ninth International Conference on Language Resources and Evaluation.* European Language Resources Association (ELRA), Reykjavik, Iceland. 2049–2056.

POLIQUEN, BRUNO – KIMLER, MARCO – STEINBERGER, RALF – IGNAT, CAMELIA – OELLINGER, TAMARA – BLACKLER, KEN – FUART, FLAVIO – ZAGHOUANI, WAJDA – WIDIGER, ANNA –FORSLUND, ANN-CHARLOTTE – BEST, CLIVE 2006. Geocoding multilingual texts: Recognition, Disambiguation and Visualisation. In: CALZOLARI, NICOLETTA – CHOUKRI, KHALID – MAEGAARD, BENTE – MARIANI, JOSEPH – ODIJK, JAN – TAPIAS, DANIEL eds., *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'20006 Genoa, Italy, 24-26 May 2006).* 53–58.

SIMON ESZTER 2013. *Approaches to Hungarian Named Entity Recognition.* Doktori (PhD) értekezés. Budapesti Műszaki Egyetem, Budapest. http://www.omikk.bme.hu/collections/phd/Termeszettudomanyi_Kar/2013/Simon_Eszter/ertekezes.pdf (2015. 12. 01.)

STEINBERGER, RALF – POLIQUEN, BRUNO – VAN DER GOOT, ERIK 2009. An Introduction to the Europe Media Monitor Family of Applications. In: GEY, FREDRIC – KANDO, NORIKO – KARGLEN, JUSSI eds., *Information Access in a Multilingual World. Proceedings of the SIGIR 2009 Workshop (SIGIR-CLIR'2009).* Boston, USA. 1–8.

STEINBERGER, RALF – EHRMANN, MAUD – MOHAMED, EBRAHIM – STEINBERGER, JOSEF – TURCHI, MARCO 2013. Multilingual Media Monitoring and Text Analysis – Challenges for Highly Inflected Languages. In: HABERNAL, IVAN – MATOUŠEK, VÁCLAV eds., *Text, Speech and Dialogue. 16th International Conference, TSD 2013, Pilsen, Czech Republic, September 1-5, 2013. Proceedings.* Lecture Notes in Artificial Intelligence 8082. Springer-Verlag, Berlin–Heidelberg. 22–33.

PAJZS JÚLIA
MTA Nyelvtudományi Intézet

JÚLIA PAJZS, Recognition of proper names in the development of the Hungarian module of the European Media Monitor

The European Media Monitor (http://emm.newsbrief.eu) developed by the European Joint Research Centre collects and categorizes news items automatically from several thousands of news sites all over the world, 24 hours a day, updating every 10 minutes, using the tools and techniques

of language technology. The full range of the service has been fitted for use in Hungarian. News items of international relevance are available in all languages involved in the project. This paper presents the attempts and achievements concerning the recognition of Hungarian proper names and the treatment of their suffixed forms within the European Media Monitor system.