# Szótövező eljárások hatása magyar szövegek automatikus kategorizálásánál

Tikk Domonkos[1], Törcsvári Attila[2], Biró György[3], Bánsághi Zoltán[1]

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem
Távközlési és Médiainformatikai Tanszék,
H-1117 Budapest, Magyar Tudósok krt. 2.
tikk@tmit.bme.hu
[2] Arcanum Development Ltd.
H-1117 Budapest, Baranyai u. 10. I/1
attila@arcanum.com
[3] TextMiner Bt.
H-1029 Budapest, Gyulai P. u. 37.
george.biro@gmail.com

**Kivonat:** Cikkünkben a szövegbányászat területén jellemzően alkalmazott vektortér-modell reprezentáció egyik fontos kérdését, a dimenzióredukciót tárgyaljuk. Ezen belül különböző szótövező eljárások hatását vizsgáljuk több szempontból. Egyrészt azt tekintjük át, milyen összefüggés van az alkalmazott szótövező és a szótár mérete között. Másrészt az egyik szövegbányászati alapfeladat, az osztályozás esetén azt tanulmányozzuk, hogy az egyes szótövezők alkalmazása milyen minőségi következménnyel jár. A vizsgálat során a HunStem szótövezőt, a szópárlista alapú szótövezőt, és egy általunk javasolt ún. óvatos szótövező eljárást hasonlítunk össze. Tesztjeink során a HITEC automatikus osztályozó programcsomagot használtuk.

## 1 Bevezetés

Szövegbányászati (text mining) és szöveges adatokon végzett információvisszakeresési (information retrieval) módszereknél leggyakrabban *vektortér-modellt* használnak a szövegek reprezentációjára, ahol a vektortér dimenziója megegyezik a vizsgált korpuszban előforduló különböző terminusok (szavak, kifejezések, n-grammok) számával, azaz a *szótár* méretével. Ez már kis méretű, azaz néhány megabájtos korpuszok esetén is igen nagyméretű lehet (ha csak szavak szerepelnek a szótárban akkor is lehet akár 100.000-es nagyságrendű), és a korpusz méretével  igaz csökkenő mértékben  tovább növekszik. Különösen igaz ez a magyar nyelvű korpuszokra, hiszen a nyelv toldalékoló jellege miatt egy terminus igen sok (akár több tucat) formában fordulhat elő. A nagyméretű szótár mind a tárigény, mind az előfeldolgozás és üzemszerű működés időigénye szempontjából hátrányos, ezért a szótár méretének csökkentése nagy jelentőségű feladat, amit minta-felismerési terminológiával a dimenzió redukálásának is neveznek.

A dimenzió redukálásának egyik leghatékonyabb módja szótövező alkalmazása, azaz amikor valamely szó toldalékolt (pontosabban többnyire csak a ragozott) előfordulásait a szótővel mint kanonikus alakkal helyettesítünk a reprezentációban. Angol nyelvű szövegek esetén ez az eljárás általában 50-65%-kal csökkenti a szótár méretét.

Cikkünkben egy igen gyakori szövegbányászati alkalmazás, a szövegosztályozás esetén vizsgáljuk meg a szótövezés hatását a szótár méretére és az osztályozás hatékonyságára vonatkozóan. A munkánk során a HITEC hierarchikus szövegosztályozót[82] és az [origo]-ról letöltött mintegy 18 ezer dokumentumot, valamint ezen portál kategóriarendszerét használtuk a tesztjeink elvégzésére.

## 2 Szótövező eljárások

Az általunk kifejlesztett HITEC szövegosztályozó programcsomag lehetőséget nyújt különböző nyelvű szövegek kezelésére és tetszőleges szótövező eljárás integrálására. A különböző nyelvek paramétereit (pl. karakterkészlet, kisbetű-nagybetű párosítás, funkciószavak listája, opcionálisan szótár megadása) egy XML formátumú nyelvdefiníciós állományban lehet megadni. Ugyanitt lehetőség van a szótövező szabályainak, illetve kivételeknek megadására is, de valamely komplett szótövező eljárást külső függvényként meghívva is aktiválni lehet a programból.

Tesztjeink során a szótövezés nélküli feldolgozást az alábbi szótövezők segítségével végzett feldolgozással hasonlítottuk össze:

- Hunstem – a Hunmorph statisztikai alapú ingyenesen elérhető programcsomag szótövező eljárása;
- Timid stemmer – egy általunk fejlesztett néhány szabály és egy korpuszfüggő *lexikon*[83] segítségével bármely nyelvre adaptálható „óvatos” szótövező algoritmus (részletes leírást ld. a 2.1 szakaszban);
- Szópárlista alapú szótövező – adott egy szavak ragozott és kanonikus alakjainak párjait tartalmazó lista, ennek segítségével egy egyszerű eljárás a ragozott alakot kicseréli a kanonikus alakra. A módszer hátránya, hogy a nagyméretű listát a memóriában kell tárolni a feldolgozás során.

Sajnos az összehasonlítás során anyagi okok miatt nem volt lehetőségünk a hazai piacon jelenlévő másik magyar termék, a Morphologic Kft. által fejlesztett HelyesLEM[84] szoftver kipróbálására.

### 2.1 Az óvatos szótövező

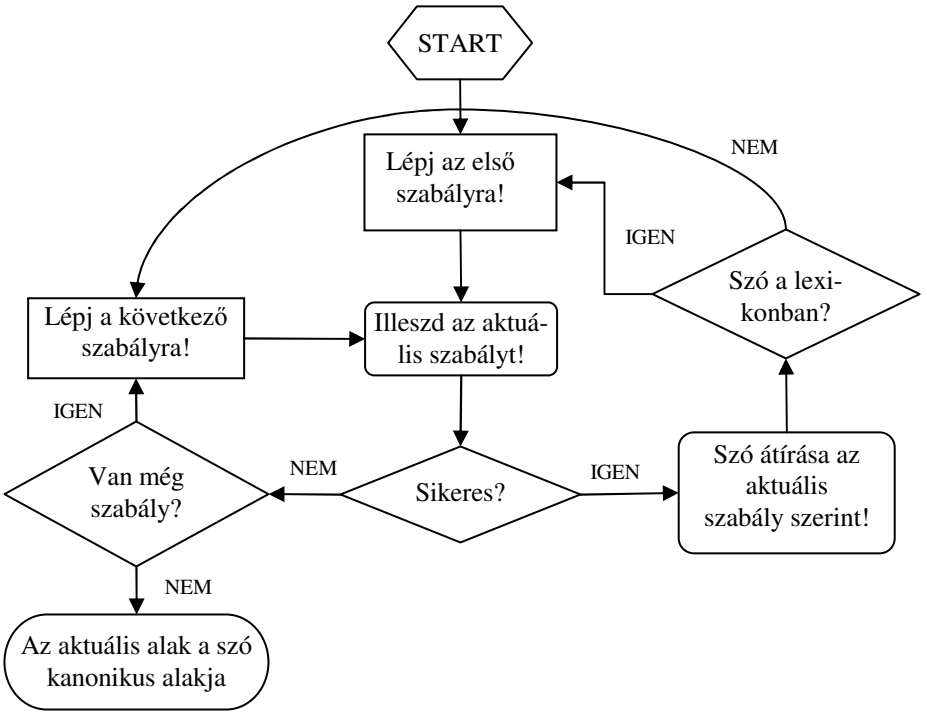
Az óvatos szótövező (timid stemmer) működési elve a következő. Legyen adott egy kiindulási lexikon, és a toldalékok levágását megadó átírószabályok sorozata. Az eljárás sorrendben illeszteni próbálja az átírószabályokat a vizsgált szóra, és amennyiben sikerül, megvizsgálja, hogy az így kapott átírt szó szerepel-e a lexikonban. Amennyiben igen, alkalmazzuk az átírószabályt és az így kapott szóra rekurzívan

[82] http://www.textminer.hu

[83] Ezt az elnevezést használjuk, hogy ne keveredjen a korpusz szavaiból álló szótár fogalommal.

[84] http://morphologic.hu/h_hlem.htm

próbáljuk a lehető legrövidebb kanonikus alakot megtalálni. Amennyiben nem, további illesztésekkel kísérletezik. Az átírószabályok sorrendje egyben prioritást is jelent. Ha az illesztés egyik lehetséges átírószabály esetén sem sikeres, akkor a szó az eredeti alakjában bekerül a lexikonba (ld. 1. ábra).




**1. Ábra.** Az óvatos szótövező folyamatábrája

Ezen a módon a szöveg maga gazdagítja a lexikont, lehetőséget nyújtva speciális szókészletű szövegek pontosabb szótövezésére. Az átírószabály lehet rag, illetve toldaléklevágó (postfix) szabály, illetve prefixátíró szabály is (pl. igekötők, és a felsőfok jelének a detektálására), valamint a hajlító nyelvek esetén szükséges reguláris kifejezések alkalmazása is. A lexikonban lehetőség van helyettesítő jelek (wildcard) használatára, valamint kivételek és szinonimák megadására is (ami egy alternatív lehetőség a dimenzió redukcióra).

Az óvatos szótövezőnek előnye, hogy nincs hozzá feltétlenül szükség kiinduló lexikonra, hiszen azt a korpusz alapján is létre tudja hozni. Természetesen egy átfogó kiinduló lexikon (illetve egyéb nyelvspecifikus adatok, pl. elhagyandó szavak listájának megadása) növeli a szótövezés hatékonyságát. További előnye, hogy némi nyelvismerettel bárki próbálkozhat átírószabályok megadására, amire a HITEC környezetében egyszerű XML elemek megadásával lehetőség van. Ebből következik, hogy bármely nyelvre könnyen adaptálható.

### 2.2 Szótövező eljárások összehasonlítása

Az alábbi táblázatban röviden összehasonlítjuk a vizsgált szótövező eljárásokat.

| Szótövező | Lexikon bővíthetősége | Szabályok bővíthetősége | Automatikus lexikonépítés | Adaptálás más nyelvekre | Átlagos hatékonyság | Futási idő | Memóriaigény |
|---|---|---|---|---|---|---|---|
| HunStem | Nehéz | Nehéz | Nincs | Nem | Kiváló | Kicsi | Kicsi |
| Óvatos | Könnyű | Könnyű | Van | Igen | Jó | Kicsi | Közepes |
| Szópárlista | Nehézkes | Nincs | Nincs | Igen | Gyenge | Min. | Nagy |

Természetesen a leghatékonyabb a magyar nyelvre specializált HunStem eljárás, és külön ki kell emelni, hogy itt a felhasználónak semmilyen nyelvészeti képzettséggel nem kell rendelkeznie. Hátránya azonban, hogy nehézkesen módosítható, hiszen nincsen erre alkalmas felülete, és ezért nehéz valamely szakterület korpuszára, pl. genetikai, vagy kémiai szövegekre alkalmazni. A másik két módszernek fő előnye a rugalmasság, viszont a felhasználónak nagyobb energiát kell befektetnie egy használható verzió létrehozásába. A szótövezők memóriaigénye fordítottan arányos az egyszerűségükkel: a legkisebb tárigénnyel a Hunstem módszer bír. A másik két eljárás esetén a lexikon számottevő memóriatöbbletet jelent, főleg a szópárlista, amelynek mérete 64 MB, több, mint 2,7 millió szópárt tartalmaz, és ami a program futása során kb. 350 MB memóriát igényel.

### 2.3 HITEC dimenziócsökkentő paraméterei

A HITEC két alapvető dimenziócsökkentő paraméterrel rendelkezik, amely az osztályozási feladatra vonatkozó alábbi intuitív megállapításokat használják ki

- A nagyon alacsony gyakoriságú szavak elhagyhatók, mert nem befolyásolják jelentékenyen az osztályozás minőségét. Küszöbérték: $\min(d_1)$
- Az olyan szavak, amely a dokumentumok (és így persze a kategóriák nagy részében előfordul) nem bírnak megkülönböztető jelleggel a kategóriák között, és így elhagyhatók. Küszöbérték: $\max(d_2)$

A két paraméter optimális értéke természetesen függ a korpusz méretétől és a dokumentumok sokféleségétől, azonban már viszonylag kicsiny $d_1$ érték esetén jelentősen csökken a szótár mérete.

## 3 Eredmények

A cikkünkben vizsgált korpusz 130219 az [origo] portálról letöltött dokumentumot tartalmaz. A lemmatizálás után 215423 különböző szót tartalmaz. A szópárlista alapú szótövező ezen szavaknak csak 55%-át ismeri, így amennyiben az ismeretlen szava-

kat változatlan formában hagyjuk, akkor 122027 szó, amennyiben ezeket elhagyjuk, akkor csak 40034 marad a szótárban.

Vizsgálatainkban megmutatjuk, hogy magyar nyelvű szövegek esetén bármely szótövező alkalmazása jelentősen csökkenti a szótár méretét, ez akár 80% feletti eredményt is adhat.

Az osztályozás hatékonysága azonban jóval kisebb mértékben függ a szótár méretétől: a legegyszerűbb szótövező is közel olyan hatékony, mint a Hunstem eljárás.

## 4 Köszönetnyilvánítás

Tikk Domonkost az MTA Bolyai János kutatói ösztöndíja támogatta.