Magyar nyelvű néprajzi keresőrendszer

Zsibrita János[1], Vincze Veronika[2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
zsibrita@inf.u-szeged.hu
[2] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
vinczev@inf.u-szeged.hu

Kivonat: A cikkben bemutatjuk Java-alapú keresőrendszerünket, mely különféle néprajzi szövegekben – hiedelmekben, táltosszövegekben és népmesékben – egész mondatos kereséseket tesz lehetővé. A rendszer azokat a dokumentumokat adja vissza, ahol a keresett ige és annak vonzatai a keresőkifejezésben megadott nyelvtani viszonyban állnak egymással. A rendszer alapjait a magyarlanc morfológiai és szintaktikai elemző moduljai jelentik. A kereső a teljes egyezések mellett részlegesen egyező találatokat is képes visszaadni, illetve a találatok grafikus megjelenítésére is van mód.

1 Bevezetés

A MASZEKER projekt keretében az angol nyelvi szabadalmi keresőrendszer mellett [1] egy magyar nyelvű néprajzi szövegeken működő keresőrendszer is elkészült. A kereső célja, hogy különféle néprajzi dokumentumokban egész mondatos kereséseket hajtson végre, azaz olyan dokumentumokat ad vissza, ahol a keresett ige és vonzatai a keresőkifejezésben megadott viszonyban állnak egymással. A keresőrendszer teljes egészében Javában implementált, így platformfüggetlenül használható.

2 A keresőrendszer

A keresőrendszer alapjait a magyarlanc morfológiai és szintaktikai (dependencia)elemző [3, 4] jelenti, amely meghatározza a szövegben levő szavak szófaját és a köztük levő nyelvtani kapcsolatokat. A keresés hátteréül szolgáló adatbázis magyar nyelvű hiedelmeket, táltosszövegeket, illetve meséket tartalmaz, összesen kb. 1,4 millió szövegszóból áll [2]. A néprajzi szövegek hatékony nyelvi elemzéséhez szükségesnek bizonyult a népies, illetve régies helyesírású szavak mai helyesírás szerinti átírása, így első lépésben ezek cseréje történt meg egy, a magyarlancba integrált szótáralapú hibajavító modul segítségével.

A keresés során a keresőmondatot először dependenciaelemzésnek vetjük alá, majd a megtalált grammatikai relációknak megfelelő illeszkedéseket keresünk a szövegekben: a keresőkifejezés igéjének összes előfordulását megkeressük, majd megnézzük, hogy az adott mondatokban levő igei bővítmények lemmája egyezik-e a

keresőkifejezésben szereplő bővítmények lemmájával, illetve hogy ugyanolyan grammatikai reláció van-e köztük (pl. mindkét esetben az ige tárgyáról van-e szó). Amennyiben igen, teljes értékű találatként adja vissza a rendszer az adott mondatot, illetve dokumentumot. Ha csak részleges egyezést tapasztalunk, például az ige egyik bővítménye egyezik, de a másik eltér, akkor azt részleges találatként jeleníti meg a rendszer. Lehetőség nyílik arra is, hogy csak az ige egyezését vizsgáljuk. A keresés során választható, mely részkorpusz szövegeiben kívánunk keresni, illetve a találatok szintaktikai reprezentációjának grafikus megjelenítésére is van lehetőség.

2.1 A keresés során használt korpusz jellemzői

A demonstráció egy magyar néprajzi korpuszon történik, aminek a konszolidálása, azaz a benne szereplő népies írásképű szavaknak a ma szokásos alakra alakítása (az eredeti íráskép megtartása mellett) már korábban megtörtént. A hiedelem- és a táltosszövegek a Néprajzi Múzeumnak a történelmi Magyarországról származó gyűjtéséből származnak, korabeli feljegyzések alapján gépelték be a kutatók. A mese korpusz néhány ingyenes internetes forrásról történő gyűjtés eredménye.

- Állatmesék (124 dokumentum)
- Formulamesék (20 dokumentum)
- Hazugságmesék (12 dokumentum)
- Legendamesék (55 dokumentum)
- Novellamesék (136 dokumentum)
- Rászedett ördög mesék (5 dokumentum)
- Rátótiádák (1 dokumentum)
- Tréfás mesék (11 dokumentum)
- Trufák és anekdoták (23 dokumentum)
- Tündérmesék (124 dokumentum)

A korpusz fontosabb adatai az 1. táblázatban láthatók.

1. táblázat: A néprajzi korpusz mérete.

Szövegtípus	Szövegek száma	Szavak száma
Népi hiedelem	2704	65807
Táltosszövegek	432	44021
Népmesék	505	633047
Összesen	**3641**	**742875**

2.2 A keresőkifejezés kialakítása

A keresés megszorított nyelvezetű keresőkifejezések alapján történik. A keresőkifejezés egy **egy tagmondatból** álló, **egy igét** tartalmazó (egyszerű) mondat. Az ige bővítményeként különféle **főnevek** szerepelhetnek különféle esetragokkal. A

határozószavak használata nem megengedett. **Tagadást** és **modalitást** jelző elemek használatát sem engedjük meg.

Néhány jólformált keresőkifejezés:

- *Foggal születik a táltos.*
- *A róka kergeti a nyulat.*
- *A lány körbefutja a házat.*

Néhány rosszulformált keresőkifejezés:

- *A vörös róka kergette tegnap a nyulat.*
- *A lány hirtelen előveszi és megeszi az almát.*
- *A róka nem eszi meg a nyulat.*
- *A róka megeheti a nyulat.*

Alárendelő mellékmondatok, illetve **mellérendelő tagmondatok** használata sem megengedett, ilyenkor több egymás utáni keresőmondatot kell alkalmazni.

- *A lány körbefutja a házat, és megeszi az almát. -> A lány körbefutja a házat. A lány megeszi az almát.*

Természetesen az alany pontos meghatározása nélkül is lehetséges keresni, ilyenkor csak az ige egyéb vonzatait tüntetjük fel a keresőkifejezésben:

- *Foggal születik.*
- *Bikával küzd.*

2.3 A keresőrendszer korlátai

A magyar nyelv grammatikai sajátosságaiból adódóan azonban problémát jelentenek a névmási referenciák, illetőleg az olyan mondatok, ahol a bővítményeket nem fejezzük ki külön szóval. Lásd az alábbi szövegrészletet:

> *Volt egyszer egy király$_i$, aki$_i$ olyan gyönyörű templomot építtetett, hogy közel s távolban nem találta senki párját. Egy szép napon azután messze távolból jött egy vándor$_j$, és (ő$_j$) hosszan bámulta a csodaszép épületet. A király$_i$ odament hozzá$_j$, és (ő$_i$) megkérdezte tőle$_j$, hogy tetszik neki$_j$ a templom.*

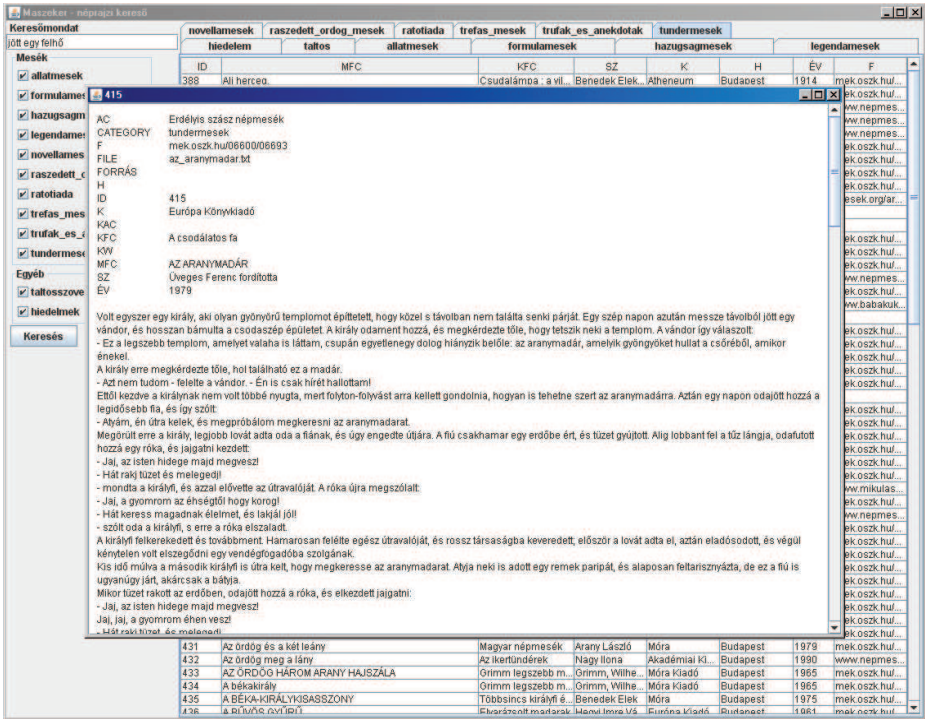
A szövegben azonos indexszel vannak jelölve az azonos egyedre utaló szavak, a zárójelbe tett névmások pedig az eredeti szövegben nem szerepelnek. Azonban legjobb tudomásunk szerint a magyar nyelvre nem áll rendelkezésre olyan automatikus elemző, amely az ehhez hasonló eseteket automatikusan azonosítaná, így jelenleg ez a

szöveg nem jelenik meg találatként az „a király odamegy a vándorhoz” keresőkifejezésre.

3 Az eredmény bemutatása

A keresés gomb megnyomásával elindul a keresés és az illesztés algoritmusa. A keresési algoritmus lefutása után megjelenik a keresőkifejezés elemzett fastruktúrája. Ezek után pedig a keresésnek megfelelő dokumentumokból készített találati lista.

Ha bármely találati listában szereplő dokumentum teljes tartalmát meg szeretnénk tekinteni, elég a dokumentum sorára kattintanunk. Ekkor egy új ablak nyílik meg, a teljes dokumentummal.

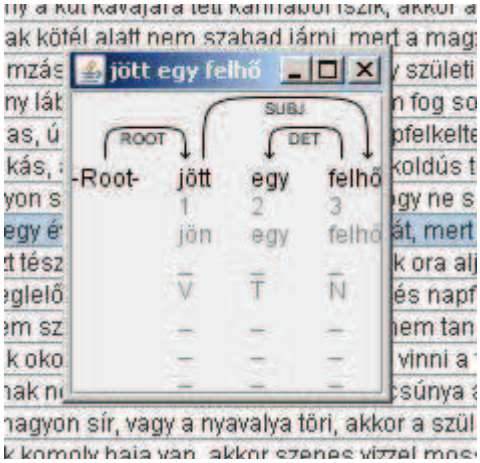


1. ábra. A találati lista egy megnyitott dokumentummal.

A megjelenő új ablakban a dokumentumra jellemző egyéb metainformációk is megjelennek, azaz annak kategóriája, az internetes forrás elérhetősége, a fájl neve, azonosítója, a kötet címe, kiadási helye és éve, a gyűjtő vagy a kötetszerkesztő neve:

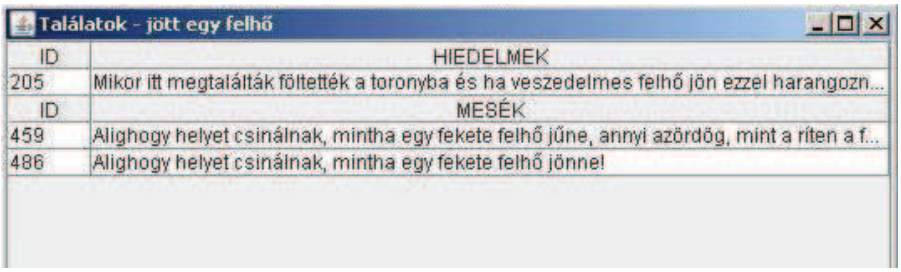
AC
CATEGORY tundermesek
F
www.nepmese.hu/index.php?option=com_mtree&Itemid=26
FILE bruncik_kiralyfi.txt
FORRÁS
H Budapest
ID 459
K Akadémiai Kiadó
KAC
KFC Rózsafiú és Tulipánleány
KW
MFC Bruncik királyfi
SZ Kovács Ágnes szerk.
ÉV 1987

Az elemzett keresőkifejezés megjelenítésére is van lehetőség, l.2. ábra.



2. ábra. Egy elemzett keresőkifejezés megjelenítve.

A keresés eredménye egy új ablakban jelenik meg, dokumentumcsoportok szerint rendezve, ahogyan az a 3. ábrán is látszik. A találati lista minden sora egy-egy dokumentumot reprezentál. Minden sor tartalmazza az adott dokumentum keresőkifejezésre illeszkedő mondatát, vagyis a releváns szövegrészt.



3. ábra. A „jött egy felhő” keresőkifejezésre illeszkedő dokumentumok találati listája.

A találati lista egy tetszőleges elemére kattintva megjeleníthető a releváns mondat elemzése (l. 4. ábra). Így ellenőrizhető, hogy az algoritmus mi alapján végezte el az illesztést.

4. ábra. A találati lista egy elemének szintaktikai elemzése.

A néprajzi keresőrendszer egy népmesékben való keresést lehetővé tevő verziója kutatási célokra nyilvánosan elérhető a http://maszeker.huminf.u-szeged.hu honlapon.

Köszönetnyilvánítás

A kutatás – részben – a MASZEKER kódnevű projekt keretében a Nemzeti Fejlesztési Ügynökség támogatásával, illetve a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió és az Európai Szociális Alap társfinanszírozása mellett valósult meg.

Hivatkozások

1. Szőts M., Csirik J., Gergely T., Karvalics L.: MASZEKER: projekt szemantikus keresőtechnológia kidolgozására. In: Tanács A., Vincze V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Tudományegyetem, Szeged (2010) 159–167
2. Szőts, M., Darányi, S., Alexin, Z., Vincze, V., Almási, A.: Semantic processing of a Hungarian ethnographic corpus. In: Proceedings of the First International AMICUS Workshop on Automated Motif Discovery in Cultural Heritage and Scientific Communication Texts. Bécs, Ausztria (2010) 112–115
3. Zsibrita J., Vincze V., Farkas R.: Ismeretlen kifejezések és a szófaji egyértelműsítés. In: Tanács A., Vincze V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 275–283
4. Zsibrita J., Vincze V., Farkas R.: magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés. In: Tanács A., Vincze V. (szerk.): IX. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2013) 368–374