MTA Nyelvtudományi Intézet
IHM-ITEM 165/2003

Intelligens többnyelvű dokumentumkezelés EUROVOC rendszerben

Záró beszámoló

Budapest, 2005. január 7.

Tartalomjegyzék

Ábrák jegyzéke

1. fejezet

Dokumentumkezelés EUROVOC rendszerben

1.1. Bevezetés

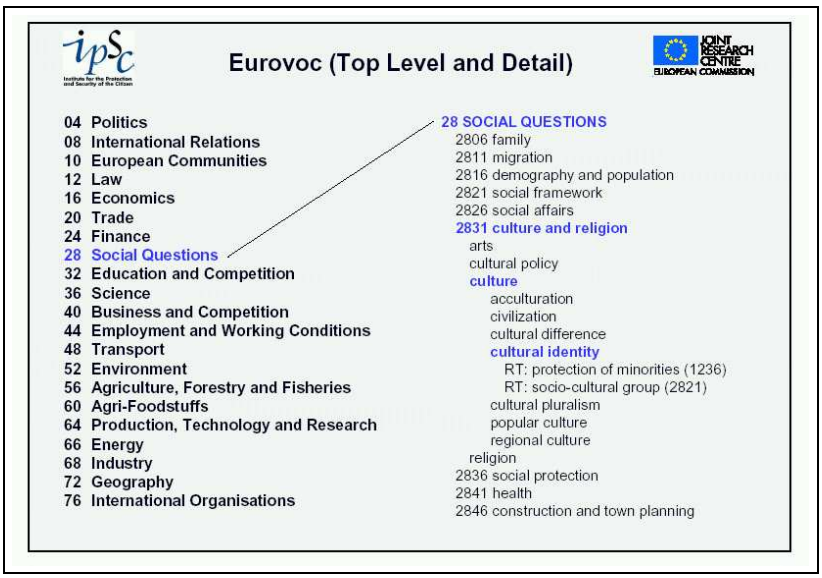
A szakmai záró beszámoló az alábbiak szerint épül fel. Az első fejezet a projektum általános leírását tartalmazza, a tudományos és technológiai háttér, a célkitűzések és javasolt megoldások tömör, lényegre szorítkozó jellemzésével. Ez után a további fejezetek a munkaszakaszokban elvégzett tevékenységek leírását tartalmazzák, az egyes feladatok elméleti hátterére és indoklására csupán akkor térünk ki, ha azok az általános leírásban nem kerültek tárgyalásra. Az egyes munkaszakaszokhoz kapcsolódó eredmények a projektum honlapján* férhetőek hozzá.

A projekt célja egy többnyelvű dokumentumkezelő rendszer kifejlesztése, amely a dokumentumok tematikus besorolását az EU által kidolgozott, és rendszeresített EUROVOC osztályozási rendszer kategóriái szerint végzi el. Az EUROVOC fogalmi osztályozási rendszer (tezaurusz) kategóriái egyértelmű megfelelésben állnak a különböző nyelvek között, ezért az EUROVOC kategóriáival indexált dokumentum tartalma, az összes olyan nyelven is automatikusan ismertté válik, amelyre az EUROVOC rendszert lefordították.

*http://corpus.nytud.hu/eurovoc/eredmenyek.html

1.2. Az EUROVOC tezaurusz

A tezaurusz az európai intézmények tevékenységének gyakorlatilag minden területét lefedi. Az EUROVOC 6501 deskriptort (kulcsszót) tartalmaz, melyek maximum 8 szintű hierarchiába vannak rendezve, a legfelső szinten 21 alkategóriával, melyek alá egy szinttel lejjebb 127 ún. mikrotezaurusz tartozik. A szűkebb ill. tágabb értelmű kifejezéseket 5877 reláció köti össze, és a tezaurusz különböző részeiből 2730 reláció kapcsolja össze a valamely módon összefüggő kifejezéseket (1.1. ábra).



1.1. ábra. Az EUROVOC tezaurusz alkategóriái (Steinberger, 2000).

Az EUROVOC rendszer rendkívül nagy előnye más természetes nyelvi vagy fogalmi osztályozó rendszerekkel szemben, mint például a széles körben használt WORDNET (Miller, 1995), hogy a rendszer deskriptorai egy numerikus indexrendszer segítségével kölcsönösen és egyértelműen leképezhetők az egyik nyelvű változatról a másikra. A különböző nyelvű dokumentumok osztályozása gyakorlatilag az indexek alapján történik. Ha egy tetszőleges nyelvű szöveget sikerül indexálni EUROVOC deskriptorokkal, akkor annak besorolása nemcsak az összes olyan nyelven elérhető, amelyre lefordították az EUROVOC rendszert, de a numerikus index alapján még olyan nyelvek esetében is egyértelmű, amelyekre még nem fordították le.

Az EUROVOCban szereplő deskriptorok legtöbbje absztrakt, többszavas kifejezés, amelyeket a szövegek általában nem is tartalmaznak, ezekre a desk-

riptorokra való közvetlen keresés a dokumentumokban tehát nem vezet eredményre. Ehelyett a deskriptorokkal szemantikai vagy más kapcsolatban lévő, ún. asszociáltak listájával lehet dolgozni.

1.3. Dokumentumosztályozás EUROVOC alapon

A projektum fontos kiinduló erőforrása az Európai Bizottság isprai kutatóközpontjában (JRC-IPSC)[†] Ralf Steinberger és munkatársai által kifejlesztett olyan eljárás, amely automatikusan, tanuló korpusz alapján állít elő a deskriptorokkal asszociált, a dokumentumokra jellemző kulcsszólistákat (Steinberger, 2001). Ezek segítségével hatékonyan és a manuális osztályozáshoz képest összehasonlíthatatlanul gyorsabban lehet a dokumentumokat EUROVOC alapon kategorizálni. A technológia eredetileg angol, francia és spanyol nyelvre lett optimalizálva. minthogy azonban alapelveiben nyelvfüggetlen, természetes kiterjesztésként adódott magyar nyelvre történő adaptálása.

A dokumentumosztályozó eljárások ugyan némileg eltérő módon, de kulcsszavakat használnak. Ezek előállítására általában kétféle módszert szoktak megkülönböztetni: kulcsszókinyerést (extraction), ill. kulcsszó- hozzárendelést (assignment). A kulcsszókinyerés a szövegekben szó szerint előforduló szavakat jelöli meg kulcsszóként, míg a kulcsszó-hozzárendelés egy korlátozott szótárból (tezauruszból) választja ki a szövegekre jellemző kulcsszavakat, amelyek egyébként nem feltétlenül szerepelnek magában a szövegben. Ez utóbbi módszerhez előre indexelt tanítókorpusz szükséges, amelynek segítségével a kontrollált szótár elemei (deskriptorok) és a szövegekben ténylegesen előforduló szavak/kifejezések között teremthetünk kapcsolatot.

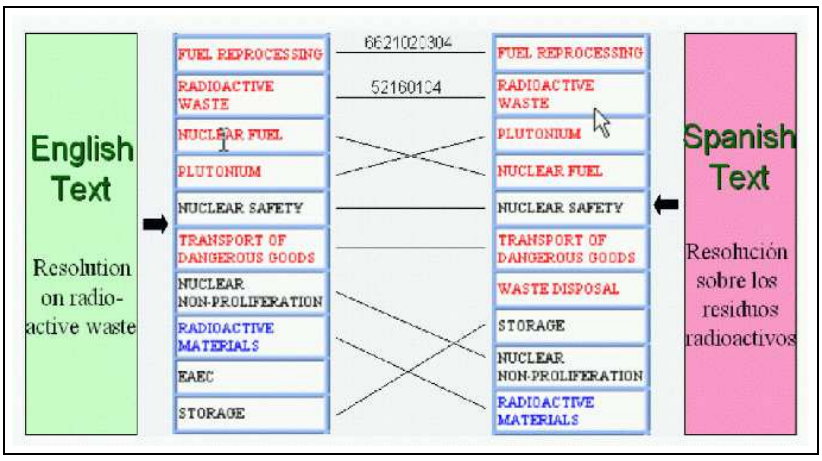
A tezauruszok használatát az motiválja, hogy hatékonyan használhatók dokumentumok osztályozására, tartalmának összegzésére és összehasonlítására. Néhány kiválasztott elemmel leírható, hogy egy-egy dokumentum milyen témakörbe tartozik, hierarchikus elrendezésük pedig lekérdezéskor lehet hasznos. A többnyelvű tezauruszok segítségével egy idegen nyelvű szöveg tartalmáról anyanyelvünkön kaphatunk információt.

Egy egynyelvű kulcsszóbesoroló eszköz statisztikai alapon választja ki egy szöveg jellemző szavait úgy, hogy az adott szöveg szavainak gyakorisági listáját összehasonlítja egy referenciakorpuszéval. Ha egy szó szignifikánsan gyakrabban fordul elő egy szövegben, mint egy „általános" szövegben (referencia-

[†] `http://www.jrc.cec.eu.int/langtech/`

korpuszban), akkor ez a szó a szövegnek egy jellemző kulcsszava lesz. A kulcsszavakhoz a szövegbeli fontosságukat jelző tartalomleíró mutató is tartozik (keyness). A felhasználók a felajánlott kulcsszavak alapján képet kaphatnak a dokumentum tartalmára vonatkozóan. Az eszköz bármely nyelvre alkalmazható, ha egy gyakorisági lista és egy lemmatizáló program rendelkezésre áll. Mivel a funkciószavak és egyéb gyakori szavak kulcsszóként nem értelmezhetők, egy ún. tiltószó (stop-word) listába felvéve kizárhatók a vizsgálat alól. A listába olyan szavak is bekerülhetnek, amelyek egyes területeken belül nem relevánsak (mert pl. minden dokumentum tartalmazza őket).

Egy ilyen statisztikai eszköz azonban nem végez semmiféle konceptuális absztrakciót, mivel csak olyan szavakat ad vissza, amelyek a szövegekben is szerepeltek. Ez a szövegek összehasonlításánál hátrányt jelenthet. Hiába hasonló ugyanis két szöveg, ha ugyanannak a fogalomnak más-más megnevezéseit (ti. szinonimákat) használja, akkor egy pusztán szavakkal operáló eljárás nem képes e hasonlóságot felfedezni. Erre a problémára megoldást jelenthet, ha a kulcsszavakat egy kontrollált szótárból vesszük, azaz a szinonimák közül mindig ugyanazt választjuk ki, bármelyik is legyen az aktuális szövegben. Ez a kontrollált szótár a jelen projektben az EUROVOC tezaurusz. Tekintve, hogy az EUROVOC rendszer egyértelmű megfeleltetéseket tartalmaz az EU összes hivatalos nyelvén, ezért használata egyben biztosítja a többnyelvű hidat a különböző nyelvű dokumentumok között (1.2. ábra).



1.2. ábra. Többnyelvű szövegek megfeleltetése EUROVOC deskriptorok alapján.

Az EUROVOC deskriptorok absztrakt fogalmakat tartalmaznak, amelyek nem feltétlenül fordulnak elő a szövegben, ezért a velük szemantikai vagy más kapcsolatban lévő, ún. asszociáltak listájával kell dolgozni. Az asszociált

terminusok olyan lexikai elemek, amelyek ténylegesen előfordulnak a szövegekben. A JRC által kidolgozott eljárás azon az elgondoláson alapul, hogy ha egy szövegben egy bizonyos deskriptorhoz sok asszociáltat találunk, akkor jó eséllyel mondhatjuk, hogy az illető deskriptor a szöveg kulcsszava. A nyelvfüggetlen módszer kézzel indexelt tanítóanyagot, valamint a fent említett statisztikai eszközt használja, és az EUROVOC deskriptorait rangsorolva, egy relevanciát mérő súllyal rendeli hozzá a szövegekhez.

Az eljárás egy kezdeti tanító fázist igényel, melynek célja, hogy minden deskriptorhoz (az adott nyelven) előálljon egy-egy asszociáltakat tartalmazó szólista, amelyben minden szóhoz egy a deskriptorhoz való asszociáltság mértékét mérő súly tartozik. Ez a súly az asszociáltak deskriptorokhoz tartozó statisztikai (és szemantikai) kapcsolódását méri. Az asszociáltak előállítása a következő módon történik. Az egyes deskriptorokhoz tartozó, előre indexelt szövegekből egyetlen ún. meta- dokumentum, tanulókorpusz készül, amin aztán lefut a fent említett egynyelvű kulcsszóbesoroló eszköz. E folyamat eredményeként előállnak az összegzett szöveg legjellemzőbb szavai, azok tehát, amelyek a legjobban asszociálhatók az adott deskriptorhoz. A fentieket minden olyan deskriptorra megismételjük, amelyekhez van elegendő (legalább 2-3 oldalas) tanítóanyag. A tanító folyamatot csak egyszer kell elvégezni minden egyes nyelvre, és az ismeretlen szövegeket kategorizáló rendszer csak az ebből kialakuló, asszociáltakat tartalmazó listát használja fel.

Új szövegek besorolásakor az eljárás az ismeretlen szöveg szógyakorisági listáját az EUROVOC deskriptorokhoz asszociált kulcsszólistájával hasonlítja össze. A szöveg gyakorisági listáját végignézve minden, az asszociáltak között szereplő szó frekvenciáját (gyakoriságát) megszorozza az asszociált (kulcs)szó súlyával, és ezt az értéket hozzáadja az asszociálthoz tartozó deskriptor pontszámához. Egy-egy szó természetesen több deskriptornak is lehet asszociáltja, de mivel ezeknek a súlya különböző, a deskriptorok más-más pontszámot kapnak. Miután a szöveg összes szavát végignézte, a rendszer a deskriptorokat pontszámuk szerint sorba rendezi, majd a kívánt nyelven megjeleníti (a fenti folyamat a dokumentum nyelvén fut le, de ha a deskriptorok az adott nyelvre egyértelműen vannak lefordítva – vagyis az EUROVOC az adott nyelven rendelkezésre áll –, az eredmény természetesen más nyelven is megjeleníthető).

A kézzel történő kulcsszókiválasztás és az automatikus módszer összehasonlítása alapján elmondható, hogy a kézi módszerrel kiválasztott deskriptorok legtöbbször szerepelnek az automatikusan kiválasztottak között, igaz nem mindig a legelsők között. Az automatikus módszer által ezeken kívül megtalált deskriptorok közül a legtöbb szintén relevánsnak tekinthető, van néhány, ami szemantikailag odaillő, de a dokumentum tartalmát nem tükrözi, kis

számban pedig nyilvánvalóan hibás deskriptort ad ki a rendszer (Steinberger, 2000). A dokumentumosztályozás hatékonysága azonban ugrásszerűen megnő, még akkor is, ha az automatikus eljárást egy manuális ellenőrző szakasz is követi.

2. fejezet

I. munkaszakasz (2004. január 4. - március 26.). Előkészítés, korpuszgyűjtés, EUROVOC és kulcsszólista fordítás

2.1. Technológiai előkészítés

A projektum első feladata a rendelkezésre álló tartalmi és technológiai források előkészítése, rendszerezése és összegzése. Ennek során egyrészt az EUROVOC és a kulcsszólisták fordítási munkálataihoz szükséges munkakörnyezet előkészítése, másrészt a JRC-féle eljárás technológiai hátterének megteremtése történt meg. Angol-magyar, francia-magyar és német-magyar gépi általános és szakszótárak lettek előkészítve a kulcsszólista fordítási feladatokra, illetve megtörtént az idegen nyelvű kulcsszólisták és forrásdokumentumok beszerzése az isprai kutatóközpontból.

2.2. EUROVOC fordítás

A munkaszakasz 2. részfeladata az EUROVOC magyar nyelvű munkaváltozatának előállítása a rendelkezésre álló számítógépes általános és szakszótárak segítségével. Az elkészült fordítás xml formátumban a honlapon hozzáférhető.

2.3. Kulcsszólisták fordítása

Ezzel gyakorlatilag párhuzamosan kezdődött a kulcsszólisták fordítása, angol, francia és német nyelvből. A kulcsszólisták közvetlenül lefordított változatára a később ismertetett egyik tanulókorpusz előállítási módszer nyersanyagaként volt szükség. A fordítási feladatoknak folyamatos háttértámogatást adott a pályázók rendelkezésére álló számítógépes általános és szakszótárak is. A kézi és gépi fordítási módszerek részletes leírása a 3.3. részben található.

2.4. Korpuszgyűjtés

A tanuló és referenciakorpusz alapanyagaként az EU jogi szövegeit tartalmazó CELEX adatbázis* szolgált. Az adatbázisból 8309, összesen mintegy 19 millió szövegszót tartalmazó magyar nyelvű dokumentum került felhasználásra. A dokumentumok legmegbízhatóbb forrása Microsoft Word formátumban volt hozzáférhető†, ezért első lépésben a hasznos szöveg kinyerése történt meg az `antiword` (`http://www.winfield.demon.nl/`) szűrőprogram segítségével. A szűrt formátumot a 2.1. ábra illusztrálja.

Példa

```
A TANÁCS HATÁROZATA
                                (1998. június 16.)
 az Európai Közösségnek a Földközi-tengeri Általános Halászati
             Bizottsághoz történő csatlakozásáról
                                (98/416/EK)


AZ EURÓPAI UNIÓ TANÁCSA,

tekintettel az Európai Közösséget létrehozó szerződésre, és
különösen  annak 43. cikkére, összefüggésben 228. cikke (2)
bekezdésének első  mondatával  és 228. cikke (3) bekezdésének
második albekezdésével,
```

2.1. ábra. A korpusz anyaga szűrt szöveges formátumban.

*`http://europa.eu.int/celex/htm/celex_hu.htm`

†A html változatok sokszor le nem fordított, angol nyelvű szöveget tartalmaztak.

3. fejezet

II. munkaszakasz (2004. március 29. - szeptember 10.). Tanulókorpusz előállítása, *deskriptor → asszociált lista* leképezés

3.1. Indexálási technológia kidolgozása

E munkaszakaszban megtörtént annak a technológiai láncnak az összeállítása, melynek segítségével a tanuló illetve referenciakorpusz automatikus nyelvi annotációja elvégezhető. A lépések részletes leírása a 3.2. részben található.

Ugyancsak ebben a részfeladatban lett kifejlesztve az a magyar nyelvre jellemző tiltószó lista, amit a kulcsszókijelölő algoritmus használ fel. A tiltószó lista kiinduló anyaga az 1720 szavas angol lista volt, amely tartalmaz funkciószavakat, pl. névmások, kötőszók, és tartalmas, az EU-szabályzatokhoz kötődő szókincset is, pl. *allow, arrangements, committee, objective, operate.* Ezen felül egy- és többszavas kifejezések szerepelnek benne (pl. *necessary to comply with this directive, forward this resolution to the commission*).

Az alaplistát az angol lista magyarra fordítása szolgáltatta. A fordítás során a magyar CELEX szövegekben folyamatosan ellenőrzésre kerültek a feltételezett magyar megfelelők. Ezután a CELEX szövegekből készült egy-, két- és háromszavas gyakorisági lista első 3,000 eleméből a nagyon gyakori szavak, a funkciószavak, illetve az EU döntéshozással kapcsolatos általános szókincs

részeinek tekinthető kifejezések kerültek még bele a magyar tiltószó listába, amely összesen 2,017 szót tartalmaz (1,265 egyszavas, 752 többszavas). A lemmatizált tiltószó listát (honlapon hozzáférhető) a 3.1. ábra illusztrálja.

Példa

```
alapelv
alaprendelet
alapszabály
alapul
alapuló
alapvető
alapvető követelmény
alapított
alap határoz meg
alap kell meghatároz
alap történő
alatt
alatti
albekezdés
albekezdés említ
albekezdés egészül ki
alkalmas
alkalmaz
alkalmazandó
alkalmazandó valamennyi
alkalmazható
alkalmazás
alkalmazási
alkalmazás részletes szabály
alkalmazás vonatkozó
```

3.1. ábra. Magyar nyelvű tiltószó lista.

3.2. Tanulókorpusz generálás. 1. módszer

Az első módszer azon az elgondoláson alapul, hogy az egymás fordításaként már létező parallel dokumentumok esetén a idegen nyelvű szöveghez hozzárendelt deskriptorok minden valószínűség szerint a magyar fordításnak is megfelelnek. Így első közelítésnek elegendő olyan dokumentumokat

felhasználni, amelyekhez az eredeti nyelven (angol) már hozzárendelték az EUROVOC deskriptorokat, és létezik magyar nyelvű fordításuk. Kézenfekvő nyersanyagként adódik a CELEX szöveganyag, amelyben a megegyező dokumentumok fordításai nyelvfüggetlen azonosító alapján kölcsönösen megfeleltethetők egymásnak. E megfelelés alapján 8304 párhuzamos (angol-magyar) dokumentumpárt lehetett felhasználni.

A magyar nyelvű szövegek hatékony osztályozásához elengedhetetlen azok automatikus nyelvi előfeldolgozása. A nyelvi annotáció első lépése magában foglalja a mondatok azonosítását (szegmentálás), az egyes szóalakok meghatározását (tokenizálás), illetve a szóalakokhoz a morfoszintaktikai jellemzők hozzárendelését (morfológiai elemzés). Ez utóbbi folyamatot a jelenleg magyar nyelvre hozzáférhető legelterjedtebb és legjobban kidolgozott elemzővel, a Morphologic HUMOR elemzőjével (Prószéky és Kis, 1999) végeztük. A morfológiai elemzés azonban önmagában még nem elegendő, mivel a szóalakok közel egyharmadához a morfológiai elemző egynél több elemzést társít. Szükség van tehát egy újabb automatikus nyelvfeldolgozó lépésre, amely kiválsztja az adott szövegkörnyezetben legvalószínűbb elemzést. Ezt az eljárást (morfoszintaktikai) egyértelműsítésnek (*part of speech tagging*) nevezik. A Nyelvtudományi Intézetben kifejlesztett módszerrel (Oravecz és Dienes, 2002) kb. 98%-os pontosságot lehet elérni, ami nemzetközi mércével mérve is nagyon jónak mondható. Az egyértelműsített és lemmatizált anyagot a 3.2. ábra illusztrálja.

Az így kapott nyelvi információ kerül a korpusz végleges XML formátumú anyagába, amely így az eredeti szöveg mellett annak szótövesített (lemmatizált) alakját is tartalmazza (3.3. és 3.4. ábra).

Ezen szövegekhez aztán az egyedi azonosító alapján közvetlenül társíthatók az angol dokumentumpárokhoz rendelt Eurovoc deskriptorok (3.5. ábra), így rendelkezésre áll egy mintegy 19 millió szavas tanulókorpusz.

3.3. Tanulókorpusz generálás. 2. módszer

A második módszer egyrészt az angol asszociált kulcsszólisták manuálisan magyarra fordított változatát használja fel, másrészt az angol, francia és német nyelvű listák gépi szótárral készült fordítására támaszkodik a következő módon. Az egyes deskriptorokhoz tartozó 3 nyelvű asszociált kulcsszavakhoz a gépi szótárak által hozzárendelt magyar megfelelőket összegyűjtjük, majd megnézzük, melyek fordulnak elő egy adott küszöbértéknél (pl. 3) gyakrabban. Ezeket aztán az eredeti listákból származó súlyértékek átlagával vesszük

Példa

```
#               [CHUNK  <DIV FROM="1">
#               (PAR    <P FROM="1">
#               (SENT   <S>
0\0     TOK     A       BOS     a\[Det]
0\0     TOK     TANÁCS          tanács\[N][NOM]
0\0     TOK     HATÁROZATA      határozat\[N][PSe3][NOM]
0\0     PUNCT   (               (\WPUNCT
0\0     TOK     1998.           1998.\DIG
0\0     TOK     június          június\[N][NOM]
0\0     TOK     16.             16.\DIG
0\0     PUNCT   )               )\WPUNCT
0\0     TOK     az              az\[Det]
0\0     TOK     Európai         Európai\[A][NOM]
0\0     TOK     Közösségnek     közösség\[N][DAT]
0\0     TOK     a               a\[Det]
0\0     TOK     Földközi-tengeri Földközi--tengeri\[A][NOM]
0\0     TOK     Általános       általános\[A][NOM]
0\0     TOK     Halászati       Halászati\[A][NOM]
0\0     TOK     Bizottsághoz    bizottság\[N][ALL]
0\0     TOK     történő         történő\[MIF][NOM]
0\0     TOK     csatlakozásáról csatlakozás\[N][PSe3][DEL]
```

3.2. ábra. Egyértelműsített szöveg.

Példa

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE celex SYSTEM "/home2/eurovoc/dtd/celex.dtd" []>
<celex>
<documents>
<document id="hu.31998D0416" sequence="0.0" celex="31998D0416">
<title type="token">
<p>
<s>
A
TANÁCS
HATÁROZATA
(
1998.
június
16.
)
az
Európai
Közösségnek
a
Földközi-tengeri
Általános
Halászati
Bizottsághoz
történő
csatlakozásáról
...
```

3.3. ábra. Referencia és tanulókorpusz végleges XML formátumban.

Példa

```
...
<title type="lemma">
<p>
<s>
a
tanács
határozat
(
1998.
június
16.
)
az
Európai
közösség
a
Földközi--tengeri
általános
Halászati
bizottság
történő
csatlakozás
...
</document>
</documents>
</celex>
```

3.4. ábra. Referencia és tanulókorpusz végleges XML formátumban (folytatás).

Példa

```
...
31998D0415  1006070304000000     EC advisory committee
31998D0415  4421010600000000     appointment of staff
31998D0416  0806031100000000     accession
31998D0416  1016010100000000     EC agreement
31998D0416  5211020401000000     Mediterranean Sea
31998D0416  5641040700000000     fishing regulations
31998D0419  0811031700000000     third country
31998D0419  2016010400000000     import
...
```

3.5. ábra. Dokumentum azonosító → deskriptorok leképezés.

fel a deskriptorhoz tartozó asszociáltak közé. Az így eredményül kapott listát illusztrálja a 3.6. ábra.

3.4. Tanulókorpusz kézi ellenőrzés

A feladat során a 3.2. részben említett 1. módszer alapján kifejlesztett anyagból véletlenszerűen 100 dokumentumot választottunk ki, és ellenőriztük, hogy az angol eredeti szövegmegfelelők alapján hozzárendelt deskriptorok mennyiben felelnek meg a magyar fordításnak. A megfelelés kisebb értelmezési különbségektől eltekintve pontos volt, így az 1. módszer alapján nagy megbízhatóságú tanulókorpusz állítható össze.

A 2. módszer szerint előállított kulcsszólistákon alapuló tanulókorpusz fejlesztés nem hozott elfogadható minőségű eredményt, a következő részben említett okok miatt, így az osztályozó rendszer jelenleg a CELEX dokumentumokból készült tanító anyagot használja fel.

3.5. Végleges leképezés kialakítása

A CELEX szövegek alapján készült korpusz 1252 féle Eurovoc deskriptor hozzárendeléséhez szolgál tanító anyagként*. Mivel a legnagyobb tanulókorpusszal rendelkező angol rendszerben 2933 deskriptor szerepel, a CELEX

*Hasonló számú deskriptort említenek a szlovén változatot fejlesztők is egy ISPRA jelentésben (honlapon elérhető).

Példa

```
DESCRIPTOR: '2026030500000000' TH: 'consumer goods'
2-ASSOC(f=2):: minta    2.61273525448177
2-ASSOC(f=2):: hibás    2.9657716033573
2-ASSOC(f=2):: hiányos  2.9657716033573
2-ASSOC(f=3):: fogyasztási cikkek       5.94808924207817
2-ASSOC(f=2):: állapot  2.38537844014719
2-ASSOC(f=2):: javítás  2.03425431943634
2-ASSOC(f=2):: megjavít 2.21078540318516
2-ASSOC(f=3):: fogyasztó        3.2405090302418
2-ASSOC(f=2):: felhasználó      3.30005660049718
2-ASSOC(f=3):: vélelem  2.77198098608905
2-ASSOC(f=2):: feltevés 2.67619533467284
2-ASSOC(f=2):: lánc     2.65347721540916
2-ASSOC(f=2):: eladó    3.38830245207565
2-ASSOC(f=2):: hiányosság       2.24347578562441
2-ASSOC(f=2):: tökéletlenség    2.24347578562441
2-ASSOC(f=2):: hiány    2.24347578562441
2-ASSOC(f=3):: hiba     2.19143775717861
2-ASSOC(f=2):: kezes    3.56475434222067
2-ASSOC(f=2):: áruforgalom      2.15070367256642
2-ASSOC(f=2):: jótállásra kötelezett    3.56475434222067
```

3.6. ábra. Gépi fordítással indukált asszociált lista.

szövegekben nem használatos 1681 deskriptorra a 3.3. részben leírt módszert lehet használni. Ennek a módszernek az eredményességét vizsgálva azonban kiderült, hogy azokra a deskriptorokra kapott asszociált lista esetében, amelyekre létezett a CELEX tanulókorpuszból kinyert szólista is, a 2. módszer alapján kapott lista igen kis (kb. 10%) fedésben van csak a CELEX szövegekből a szabályos eljárással generált listával. Ezért a 2. módszeren alapuló asszociált listák alapján hozzárendelt deskriptorok nem tekinthetők elegendően megbízhatónak. Jelenleg tehát, amíg további nagy mennyiségű, ellenőrzött tanulókorpusz nem áll rendelkezésre[†], a mostani rendszer elvárható megbízhatósággal (részleteket lásd 4.2. rész) 1252 deskriptort képes kezelni. A végleges leképezésben (lásd 3.7. ábra illetve honlap) mindazonáltal mind a CELEX korpuszból generált (0-ás prefix), mind a kézi (1-es prefix), mind gépi (2-es prefix) fordításból származó asszociáltak szerepelnek, a teljes listából a kívánt típusú adatok egyszerű szűréssel kinyerhetők.

Példa

```
HU-DESCRIPTOR: '5211020603050000' TH: 'North Sea'
0-ASSOC:: északi        2.55700134888217
0-ASSOC:: heringállomány        2.30988423261835
0-ASSOC:: heringfogás  2.64545316349326
0-ASSOC:: tenger      2.39156880514323
1-ASSOC(e=allowable catch):: megengedett zsákmány  4.06216501330961
1-ASSOC(e=by catch):: elfogva   2.53009363814902
1-ASSOC(e=capture):: zsákmányolás       2.31450848379488
1-ASSOC(e=catch in):: elkap     2.9078628479611
1-ASSOC(e=common fishery policy):: közös halászati érdek
                                                  2.02527862830344
1-ASSOC(e=fish):: hal   2.49710150262343
...
2-ASSOC(f=2):: hering   2.15469688622923
2-ASSOC(f=2):: spratt   2.30497862630778
2-ASSOC(f=2):: tenger   4.61818336667367
2-ASSOC(f=2):: befogás  2.60725367261133
2-ASSOC(f=2):: ivadék   2.5444123759931
```

3.7. ábra. Teljes deskriptor → asszociált lista leképezés.

[†]Ennek előállítása messze túlmutat a jelen projektum vállalásain és keretein.

4. fejezet

III. munkaszakasz (2004. szeptember 13. - december 31.). Többnyelvű dokumentumkezelő rendszer implementálása

4.1. Hálózati felület fejlesztése

Ebben a szakaszban történt annak a hálózati felületnek a kifejlesztése, amely a dokumentumbesoroló rendszerhez felhasználói hozzáférést biztosít (lásd honlap). A felületet Apache web szerver, Perl modulokat használó CGI programok és PostgreSQL adatbázis kezelő rendszer szolgálja ki.

4.2. Tesztelés

A tesztelés folyamata a következő lépéseket tartalmazta:

- nyelvi előfeldolgozás
- tanító fázis mintegy 7500 CELEX dokumentum alapján
- automatikus deskriptor hozzárendelés ellenőrzése 179 tesztdokumentum alapján

A részletes eredmények, melyek más nyelvekre kapott eredményekkel összevetve hasonlóan jónak mondhatók, a csatolt kiértékelő jelentésben találhatók

(honlapon szintén hozzáférhető).

4.3. Dokumentumbesoroló eszköz implementálása

Az utolsó részfeladatban a teljes rendszer összeépítése volt a feladat. A dokumentumosztályozás során az első lépés a szövegek nyelvi előfeldolgozása (morfológiai elemzés, egyértelműsítés és szótövesítés). Ezt a folyamatot a `http://corpus.nytud.hu/postag/` oldalon található demonstrációs felület illusztrálja. Az előfeldolgozás kimenete (az egyértelműsített és lemmatizált szöveg) kerül a dokumentumbesoroló eszköz bemenetére, amely az asszociált kulcsszólistákat illetve az adott dokumentum jellemző szavait összehasonlítva rendel Eurovoc deskriptorokat a szöveghez. A hozzárendelés egy kiértékelő hálózati felületen (lásd honlap) ellenőrizhető.

4.4. Összefoglalás

A projektum eredményeként kifejlesztett dokumentumbesoroló eszköz prototípusnak tekintendő, amennyiben a CELEX adatbázisból származó szövegekből előállított tanulókorpuszt felhasználva mintegy 1200 deskriptort képes jelenleg kezelni nemzetközi összevetésben is jónak mondható megbízhatósággal. Miután azonban a technológia részletesen kidolgozott, és eredményességét az elvégzett kiértékelés egyértelműen igazolja, a rendszer *fedésének* javítása nem igényel mást, mint további tanulókorpusz (hosszadalmas és jelentős manuális munkát igénylő, de mechanikus) előállítását és felhasználását.

A projektum nem elhanyagolható fontos hozadéka a jövőbeni kutató fejlesztő munka, ezen belül a dokumentumbesoroló rendszer további tesztelése és fejlesztése tekintetében a JRC-vel kialakított kiváló együttműködés, melynek során az intézet és az isprai kutatóközpont között igen hatékony, napi munkakapcsolat alakult ki.

Függelék

A. A referencia- és tanulókorpusz XML formátumához tartozó DTD

```
<!--                                                                -->
<!--                                                                -->
<!--                    Celex test dtd                              -->
<!--                                                                -->
<!--                                                                -->
<!--   $Date: 2004/04/05/ $                                         -->
<!--   For temporary use (o.cs.)                                    -->

<!--                                                                -->
<!--                   ENTITY DECLARATIONS                          -->
<!--                                                                -->
<!ENTITY rcub  "&#x007D;"> <!--/rbrace C: =right curly bracket-->
<!ENTITY lcub  "&#x007B;"> <!--/lbrace O: =left curly bracket -->
<!ENTITY laquo "&#x00AB;">    <!--=angle quotation mark, left -->
<!ENTITY raquo "&#x00BB;">  <!--=angle quotation mark, right  -->
<!ENTITY bsol  "&#x005C;"> <!--/backslash =reverse solidus    -->

<!--                  Global attributes                         -->
<!ENTITY % a.global '
          id                ID                  #IMPLIED
          n                 CDATA               #IMPLIED
          lang              IDREF               #IMPLIED'     >

<!--              Content model declarations                    -->
<!ENTITY % base.seq   '#PCDATA | w | c'                          >


<!--                                                                -->
```

```
<!--                  ELEMENT DECLARATIONS                    -->
<!--                                                          -->


<!--                  HIGH-LEVEL COMPONENTS                   -->
<!ELEMENT celex        (documents)                             >
<!ATTLIST celex             %a.global;
          type               CDATA               "text"
          version            CDATA               #IMPLIED     >


<!ELEMENT documents        (document)+                         >
<!ATTLIST documents             %a.global;                     >


<!ELEMENT document        (title+,text,lemmatised,
                           descriptors,description)           >
<!ATTLIST document           %a.global;
          sequence           CDATA               #IMPLIED
          celex              CDATA               #REQUIRED    >
<!ELEMENT title        (p)+                                    >
<!ATTLIST title            %a.global;
          type          (token | lemma )         "token"      >

<!ELEMENT text         (p)+                                    >
<!ATTLIST text             %a.global;                          >

<!ELEMENT lemmatised   (p)+                                    >
<!ATTLIST lemmatised       %a.global;                          >

<!ELEMENT descriptors  (descriptor)+                           >
<!ATTLIST descriptors      %a.global;                          >

<!ELEMENT descriptor   (%base.seq;)*                           >
<!ATTLIST descriptor       %a.global;                          >

<!ELEMENT description  (%base.seq;)*                           >
<!ATTLIST description      %a.global;                          >

<!ELEMENT p            (s)+                                    >
<!ATTLIST p                %a.global;                          >

<!ELEMENT s            (%base.seq;)*                           >
```

```
<!ATTLIST s                %a.global;                          >

<!ELEMENT w              (#PCDATA)                             >
<!ATTLIST w                %a.global;                          >

<!ELEMENT c              (#PCDATA)                             >
<!ATTLIST c                %a.global;                          >
```

Irodalomjegyzék

Miller, George. WordNet: A lexical database for English. *Communications of the ACM*, 1995, 38(11).

Oravecz, Csaba és Dienes, Péter. Efficient stochastic part of speech tagging for Hungarian. In: *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas. 2002, 710–717.

Prószéky, Gábor és Kis, Balázs. Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, Maryland, USA. 1999, 261–268.

Steinberger, Ralf. Using Thesauri for Information Extraction and for the Visualisation of Multilingual Document Collections. In: *Proceedings of the Workshop on Ontologies and Lexical Knowledge Bases*, Sozopol, Bulgaria. 2000, 130–141.

Steinberger, Ralf. Cross-lingual Keyword Assignment. In: *Proceedings of the XV II Congress of the Spanish Society for Natural Language Processing*, 2001.