A szótárkészítés támogatása párhuzamos korpuszokon végzett szóillesztéssel

Héja Enikő

MTA Nyelvtudományi Intézet, Nyelvtechnológiai Kutatócsoport
1068, Bp., Benczúr u. 33.
eheja@nytud.hu

Kivonat: Cikkünkben egy kétnyelvű szótárak készítésének gépi támogatására irányuló módszert ismertetünk. A javasolt megközelítés alapja a párhuzamos korpuszokon végzett automatikus szóillesztés. A korpuszvezérelt megközelítés, ezen belül különösen a párhuzamos korpusz használata több okból is hasznosnak bizonyult a lexikográfia számára. Ezek közül a legfontosabb, hogy – megfelelő méretű reprezentatív korpusz használatával – a javasolt megközelítés garantálja, hogy a legrelevánsabb fordítások fognak szerepelni a szótárban. További előnyt jelent, hogy az összes korpuszbeli példamondat könnyedén hozzáférhető, így a poliszém jelentések közül nagy mennyiségű természetes adat alapján választhatjuk ki a legmegfelelőbbet. A két fenti tulajdonság különösen alkalmassá teszi az általunk javasolt módszert aktív szótárak előállítására.

1 Bevezetés

A cikkben ismertetett munka az EFNIL által finanszírozott EFNILEX[1] projekt része. A projekt azt vizsgálta, hogy a nyelvtechnológiai módszerek és eszközök – különös tekintettel a párhuzamos korpuszokra – mennyiben járulhatnak hozzá a szótárkészítési folyamathoz. A szótárkészítés automatikus támogatása elsősorban a kevéssé használt nyelvek esetében bír jelentőséggel, hiszen az ilyen nyelvpárokra íródott szótárakra alacsony a kereslet, így az ilyen munkálatok finanszírozása is korlátozott. A bemutatandó munka eredeti célja egy középméretű (kb. 15,000 szócikk) litván-magyar szótár létrehozása volt. A munkafolyamat részeként tesztelési célokra a magyar-szlovén nyelvpárt is vizsgáltuk.

Jelenleg nem létezik olyan módszer, amely lehetővé tenné szótárak *teljesen* automatikus előállítását. Így egy megfelelő lefedettségű és pontosságú lexikai erőforrás előállítása mindenképpen igényel emberi utószerkesztési munkálatokat is. Ennek fényében úgy fogalmazhatjuk meg feladatunkat, hogy célja a lexikográfusok számára olyan erőforrásokat biztosítani, amelyek a lehető legjobban csökkentik a teljes értékű, emberi felhasználásra alkalmas szótárak elkészítéséhez szükséges munkát. A fenti elvárásoknak megfelelő automatikusan generált erőforrásokat protoszótáraknak fogjuk nevezni a cikk hátralevő részében.

[1] http://www.efnil.org/projects/efnilex

Az általunk javasolt módszer alapját párhuzamos korpuszokon végzett automatikus szóillesztés képezi. Bár az automatikus szóillesztést széles körben használják szótárfejlesztésre elsősorban a gépi fordítás területén, amennyire tudjuk, mostanáig ezt a megközelítést nem használták lexikográfiai projektekben emberi felhasználásra szánt szótárak készítésének támogatására.

A következő fejezetben röviden bemutatjuk a párhuzamos korpusz használatának előnyeit a szótárkészítésben. A 3. fejezetben áttekintjük a munkafolyamatot, amely három fő részből áll: a párhuzamos korpuszok létrehozásából (3.1), a protoszótárak előállításából (3.2) és a kiértékelésből (3.3). A 4. fejezetben illusztráljuk, hogy az általunk javasolt módszer jól kezeli a poliszémiát. Az utolsó szakasz összefoglalja az eredményeket és a hátralévő feladatokat.

2 A javasolt módszer előnyei és hátrányai

Napjainkban általánosan elfogadott lexikográfus körökben, hogy jó minőségű szótárakat kizárólag korpusz alapján lehet létrehozni (ld. pl. [1]). Ennek az az oka, hogy a korpusz használata jelentősen csökkenti az egyéni intuíció szerepét a szótárkészítési folyamatban.

Azonban még forrásnyelvi és célnyelvi korpusz használata esetén is vannak olyan lépések, amelyek során a szótárkészítőnek elkerülhetetlenül támaszkodnia kell az intuíciójára. Ilyen feladatok például a szótárba felveendő jelentéssel bíró nyelvi egységek (*linguistic unit*, a tovbbiakban LU) meghatározása, ezek célnyelvre való fordítása valamint annak eldöntése, hogy a lefordított LU-k közül melyeket vonják össze célnyelvi oldalon.

A költséghatékonyság mellett a párhuzamos korpusz alapú módszer másik nagy előnye, hogy választ ad arra a kérdésre, hogy hogyan csökkenthető tovább az intuíció szerepe a lexikográfiában. Ebben az esetben az LU-kat nem a lexikográfusok nyerik ki a korpuszból, hanem a statisztikai szóillesztő algoritmus. Így nem az emberi intuíció határozza meg, hogy mi számít egy LU-nak, hanem a szavak kontextusa és célnyelvi fordítása. A megközelítés korpuszvezérelt (*corpus-driven*) jellege biztosítja, hogy a fordításjelöltek megállapítása során a gyakran használt fordítások nagyobb súllyal szerepeljenek. Így – ha rendelkezésünkre áll egy nagyméretű, reprezentatív korpusz – a leggyakrabban használt fordítások biztosan szerepelni fognak szótárunkban.

A megközelítés további előnye, hogy az automatikusan meghatározott fordítási valószínűség alapján a fordítási jelölteket sorba rendezhetjük aszerint, hogy mennyire valószínű fordításai a forrásnyelvi lemmának. Így korpuszadatok alapján megállapíthatjuk, hogy melyik egy lemma leggyakoribb jelentése. A módszer harmadik fontos jellemzője, hogy a fordításpárok természetes kontextusaikkal együtt jeleníthetőek meg. Az alábbi magyar-litván bejegyzésminta azt illusztrálja, hogy a korpuszból származó példamondatok alapján hogyan különíthetjük el egymástól egy szó különböző aljelentéseit.

1. táblázat: Bejegyzésminta a magyar-litván protoszótárból.

MAGYAR LEMMA	LITVÁN LEMMA	FORD. VSZ.	MAGYAR LEMMA-GYAKORISÁG	LITVÁN LEMMA-GYAKORISÁG
Születik	Gimti(-sta,-ė)	0.579	169	174
MAGYAR			**LITVÁN**	
Ő **1870-ben született**.			Jis **gimė 1870 metais**.	
De Fache mintha **erre született** volna.			Bet Fasas, regis, tiesiog **tam gimęs**.	
Úgy látszik, **szerencsétlen csillagzat alatt születtél**.			Turbūt **gimei po nelaiminga žvaigžde**.	
…, mert **ikrei születtek.**			…, nes jai **gimė dvynukai**.	
Maga úriembernek **született.**			Tu **gimei džentlemanu.**	
… hogy Buddha nem **lótuszvirágból született**?			…, kad Buda **gimė** ne **iš lotoso žiedo**?	

A javasolt módszer egyik hátránya, hogy kizárólag lemmák között hozható létre megfeleltetés, így jelenlegi formájában a többszavas kifejezések (nevek, kollokációk, igei szerkezetek) automatikus kezelésére alkalmatlan. A másik fő nehézség, mint azt rövidesen látni is fogjuk, hogy a párhuzamos korpusz összeállítása a kevéssé használt nyelvekre rendkívül időigényes feladat.

A következő részben a magyar-szlovén és magyar-litván protoszótárak elkészítését mutatjuk be.

3 A munkafolyamat

A munkafolyamat három fő szakaszból áll. Először a szükséges erőforrásokat és a szövegfeldolgozáshoz szükséges nyelvspecifikus eszközöket szereztük be (ld. 3.1). Ezt követően a szóillesztés segítségével és különböző szűrők alkalmazásával létrehoztuk a protoszótárakat (ld. 3.2). Az utolsó szakaszban kidolgoztuk a kiértékeléshez szükséges kategóriákat, majd elvégeztük a kiértékelést (3.3).

3.1 A párhuzamos korpuszok létrehozása

Erőforrások és nyelvspecifikus eszközök
Mivel a projekt célja a köznapi szókincset lefedő protoszótárak létrehozása volt, a szövegek gyűjtésekor a regényekre koncentráltunk. A projekt során felmerülő legnagyobb nehézséget a megfelelő mennyiségű, általános szókincsű erőforrás összegyűjtése okozta. Mivel a szlovén-magyar nyelvpár közötti közvetlen fordításokból nagy ráfordítással csak kevés szöveget[2] sikerült szerezni, és a litván-magyar nyelvpárra nem találtunk nagy mennyiségű közvetlen fordítást, úgy döntöttünk, hogy a litván-magyar párhuzamos korpuszt olyan szövegekből állítjuk össze, amelyeket egy harmadik nyelvről fordítottak le mindkét nyelvre. Sajnos azonban sem a szlovén, sem a litván esetében nem állnak rendelkezésre olyan digitális archívumok, mint a Digitális Irodalmi Akadémia[3] és a Magyar Elektronikus Könyvtár[4] a magyar vonatkozásában. Ezért a litván Vytautas Magnus Egyetemen található Számítógépes Nyelvészeti Központ segítségét vettük igénybe. Az intézmény a Litván Nemzeti Korpusz [9] és egy angol-litván párhuzamos korpusz [8] létrehozójaként birtokában van a projekt szempontjából szükséges erőforrásoknak és nyelvspecifikus eszközöknek.

A szótárkészítéshez szükséges szövegfeldolgozó eszközöket (tokenizáló, mondatra bontó, lemmatizáló – egyértelműsítéssel) eszközláncokba beépítve használtuk. A litván elemzést a már említett Számítógépes Nyelvészeti Központ (Vytautas Magnus Egyetem) végezte el. A szlovén szövegeket a Jožef Stefan Intézet honlapján[5] található eszközlánccal elemeztük [4]. A magyar korpusz annotálása pedig a Nyelvtudományi Intézet Nyelvtechnológiai Osztályán kifejlesztett MNSZ egyértelműsítő lánccal történt [7].

A párhuzamos korpuszok létrehozása
A mondatillesztést a *hunalign* mondatillesztővel [10] végeztük. Az illesztés bemeneteként a mondatok lemmatizált változata szerepelt, hogy a gazdag morfológiából fakadó adathiányt a lehető legkisebbre csökkentsük.

Mivel az eredeti feladat a protoszótárak előállítása és hasznosíthatóságuk vizsgálata volt, a rossz mondatillesztés esetleges hatásainak minimalizálására törekedtünk. Ezért először a szövegeket kézileg ellenőriztük, hogy kiszűrjük azokat a szövegrészeket[6], amelyeknek nincsen célnyelvi megfelelőjük. Az illesztés után a szlovén-magyar párhuzamos korpusz egy részkorpuszán a mondatpárokhoz rendelt konfidenciaértékek alapján megállapítottuk, hogy mi az a küszöbérték, amely felett nagy eséllyel már csak jó mondatillesztések vannak. A litván-magyar párhuzamos korpusz esetén is az itt megállapított értéket használtuk. Az 2. táblázatban az eredményül kapott párhuzamos korpuszok mérete szerepel.

---

[2] A szlovén televízió, számos műfordító és kiadó megkeresésével kb. egy 750.000 tokent tartalmazó korpuszt gyűjtöttünk.

[3] http://www.pim.hu/

[4] http://mek.oszk.hu/

[5] http://nl.ijs.si/jos/analyse

[6] A munka elvégzéséért köszönet illeti Mittelholcz Iványt.

2. táblázat: A párhuzamos korpuszok mérete.

LITVÁN	1,765,000 token	147.158 TU[7]
MAGYAR	2,121,000 token	147,158 TU
SZLOVÉN	733,000 token	38,574 TU
MAGYAR	666,000 token	38,574 TU

3.2 A magyar-szlovén és a magyar-litván protoszótárak létrehozása

A protoszótárak generálásának két fő szakasza volt. Első lépésben elvégeztük a szóillesztést. Erre a célra a GIZA++ szóillesztő szoftvert [6] használtuk. A GIZA++ a szóillesztés során fordításjelölteket hoz létre, úgy, hogy a célnyelvi és a forrásnyelvi lemmapárokhoz fordítási valószínűséget rendel. A protoszótárak kiindulási alapját ezek a fordításjelöltek képezték. Ezekből kellett kiszűrnünk a legjobb fordításjelölteket a lehető legtöbb helyes fordításjelölt megtartásával.

A második lépésben tehát ezt a feladatot kívántuk megoldani a magyar-szlovén eredmények egy mintájának kézi kiértékelésével[8]. A kiértékelés során három paramétert vettünk figyelembe: a GIZA++ által meghatározott *fordítási valószínűségnek* az értékét, valamint a *forrásnyelvi* és *célnyelvi jelöltek korpuszgyakoriságát*. Ez az előzetes kiértékelés két konklúzióval szolgált: egyfelől a fordításpár-jelöltben szereplő lemmák mindegyikének legalább 5-ször elő kell fordulnia ahhoz, hogy elegendő adat álljon rendelkezésre a fordítási valószínűség becsléséhez. Másfelől, a kiértékelt szópárok azt mutatják, hogy a fordítási valószínűségnek legalább 0.5-nek kell lennie. Ez alatt az érték alatt rohamosan csökken a fordításjelöltek pontossága. A fenti paramétereknek megfelelő fordításjelöltek 65%-a volt jó fordítás[9]

3. táblázat: A megfelelő fordításpár-jelöltek a teljes korpuszon.

	A megfelelő fordításjelöltek száma	A várhatóan jó fordításjelöltek száma
Magyar-szlovén	4969	3230
Magyar-litván	4025	2616

[7] *TU* (translation unit) kifejezést használjuk az illesztett egységek jelölésére, mert a *hunalign* engedélyezi a mondatok közötti egy-a többhöz megfeleltetéseket is.

[8] A szlovén-magyar szövegek gyűjtéséért és a magyar-szlovén kiértékelési munkák elvégzéséért köszönettel tartozom Sárossy Bencének.

[9] A magyar-litván esetében egy további korlátozást is bevezettünk: a fordításjelöltek közül kizártuk azokat a párokat, amelyek valamelyik tagjának gyakorisága több, mint 100-szorosa volt a másik tag gyakoriságának.

Mivel célunk nem tökéletes szótárak automatikus előállítása volt, hanem olyan protoszótárak készítése, amelyek a lehető legnagyobb mértékben segítik a lexikográfusok munkáját, jogosnak tűnik egy 65% körüli pontosságot becélozni, mivel könnyebb már meglévő, ám rossz fordításjelölteket kidobni, mint újakat felvenni a szótárba. Így ezeket a paramétereket elfogadva részletesen is kiértékeltük a magyar-litván protoszótárunkat. Ezt mutatja be a következő fejezet.

3.3 A magyar-litván protoszótár részletes kiértékelése

A magyar-litván protoszótár kiértékelését teljesen manuálisan végezték mindkét nyelvet egyaránt beszélő szakértők[10]. Az általánosan elfogadott kiértékelési eljárásokkal szemben itt elsődlegesen nem a *jó* és a *rossz* fordításjelöltek arányára voltunk kíváncsiak, hanem a *lexikográfiailag hasznos* és a *lexikográfiailag nem hasznos* fordításjelöltekére. Ezt a fajta megkülönböztetést egyrészt az olyan jó fordításjelöltek tették szükségessé, amelyek szótárkészítési szempontból irrelevánsak (elsősorban a túl specifikus tulajdonnevek). Másrészt a rossz fordításjelöltek komoly segítséget jelenthetnek a szótárkészítési munkában, elsősorban a kollokációk esetében, hiszen a kontextus alapján könnyű visszafejteni, hogy mi lett volna a helyes megfeleltetés. Az alábbiakban röviden összefoglaljuk azokat a fő kategóriákat (3.3.1), amelyeket a kiértékelés során használtunk, majd ismertetjük a kiértékelés módszertanát és az eredményeket (3.3.2).

A kiértékelés során használt fő kategóriák[11]

(1) Lexikográfiai szempontból hasznos fordításjelöltek:

a. Teljesen jó fordítások [H: ***gyümölcs*** – L: ***vaisius*** – *fruit*]

b. Részlegesen jó fordítások, ebben az esetben utószerkesztés szükséges, elsősorban az alábbiak miatt:
   i. rossz lemmatizáció
   ii. részleges/rossz illeszkedés a több szavas kifejezések esetében.

   Pl:
   1. összetett szavak [H: ***főfelügyelő*** – L: *vyriausiasis* ***inspektorius***],
   2. kollokációk [H: **bíborosi** testület – L: Kardinolų **kolegiją**]

c. Egyéb szemantikai viszony. Pl: hiperonímia [H: **lúdtoll** – L: **plunksna** (toll – madártoll, írótoll)])

[10] A magyar-litván szótár kiértékeléséért köszönet illeti Tölgyesi Beatrixot és Justina Lukaseviciute-t.

[11] A megadott példákban az automatikusan megállapított jelöltpárok félkövérrel vannak szedve.

(2) Lexikográfiai szempontból nem hasznos fordításjelöltek

a. Irreleváns szókincs (pl. gyakran előforduló tulajdonnevek [H: **Abdul** – L: **Abdulas**])
b. Rossz fordítások (általában a túl szabad fordítás miatt)

A kiértékelés eredménye

A fent meghatározott paramétereknek megfelelő[12] 4025 magyar-litván fordításjelölt közül 863 párt értékeltünk ki kézzel. Ebből 520 pár fordítási valószínűsége a [0,5, 0,7) tartományba, 380 pár fordítási valószínűsége pedig a [0,7, 1) tartományba esett. 63 pár fordítási valószínűsége volt 1. A kiértékelés eredményeit a 4. táblázat tartalmazza.

4. táblázat: A magyar-litván szótár kiértékelésének eredményei.

	Hasznos párok		**Nem hasznos párok**	
P(tr)1	OK	Utószerk.	Irreleváns	Rossz
[0.5,	52.1 %	32.9 %	2.3 %	12.7 %
	85 %		15 %	
[0.7, 1)	65.3 %	31.9 %	0.6 %	2.2 %
	97, 2 %		2,8%	
1	38 %	13 %	49 %	0 %

A 4. táblázat alapján a fordításjelöltek 85%-a hasznos a [0.5, 0.7) valószínűségi tartományba eső fordításpárok esetén. Ez az arány még jobb a [0,7, 1) intervallumba eső fordítási valószínűségek esetén (97,2%). Érdekes módon az 1 fordítási valószínűséggel rendelkező párok esetén ez az arány csupán 38%. Az irreleváns párok magas aránya (49%) azt mutatja, hogy ennek elsődleges oka, hogy a nevek hajlamosabbak 1 valószínűséggel együtt előfordulni.

4 A poliszémia kezelése a javasolt módszerrel

Mint már a cikk 2. szakaszában a megközelítés előnyei között említettük, az általunk javasolt módszerrel a korpuszból az összes releváns fordítást kinyerhetjük, ezáltal csökkentve a fordítói intuíció szerepét. Sőt, ezen felül a lehetséges fordítások rende-

[12] Mindkét lemma legalább ötször előfordul, a fordítási valószínűség legalább 0,5 és egyik lemma sem fordul elő százszor többször, mint a másik.

zésével elérhetjük, hogy a szó legvalószínűbb jelentéseit rangsoroljuk előre. Ezek alapján azt várjuk, hogy az általunk javasolt megközelítéssel hatékonyabban kezelhetjük a poliszémiát, mint a hagyományos vagy az egynyelvű korpuszokon alapuló lexikográfia.

Hogy a fenti hipotéziseket közelebbről is megvizsgáljuk, készítettünk egy litván-magyar protoszótárat is, amelyet – a teljes kiértékelés igénye nélkül – összehasonlítottunk a már meglévő litván-magyar szótárral [2].

Abból az előfeltevésből kiindulva, hogy „*erős korreláció figyelhető meg egy szó gyakorisága és szemantikai komplexitása között*” [1], csak azokat a litván lemmákat vettük figyelembe, amelyek legalább százszor előfordultak a korpuszban. Ezzel párhuzamosan a fordítási valószínűséget jelentősen csökkentettük: 0,5-ről 0,02-re. Az így meghatározott paraméterekkel 6550 fordításjelöltet kaptunk, amelyek 1759 litván lemmához tartoztak. Az 5. táblázat jól szemlélteti, hogy a javasolt módszerrel számos különböző fordítást nyerhetünk ki a korpuszból sorba rendezve aszerint, hogy a fordítás mennyire valószínű. Jól látszik továbbá az is, hogy a nagyon gyakori szavak esetében nagyon alacsony fordítási valószínűségű párok is adhatnak jó jelölteket.

5. táblázat: litván *puikus* magyar megfelelői.

LIT	**HUN**	**P(tr)**
puikus	remek	0.071
puikus	tökéletes	0.052
puikus	szép	0.048
puikus	pompás	0.035
puikus	jól	0.035
puikus	nagyszerű	0.035
puikus	finom	0.028
puikus	gyönyörű	0.02

A poliszémia ilyen módon való kezelése különösen alkalmasnak tűnik aktív (a forrásnyelvi beszélő célnyelven való megnyilatkozását segítő) szótárak készítésének támogatására. Szintén az aktív szótárak készítését segítik elő a korpuszból kinyert kontextusok, amelyek segítséget nyújthatnak a legjobb célnyelvi fordítás kiválasztásában. Ezt támasztja alá az alábbi ábra is:

1. ábra. Bejegyzések összehasonlítása.

Míg a hagyományos szótárban egy magyar fordítás található két kontextussal[13], addig az általunk készített szótárban négy magyar fordítás[14] található 75 kontextussal.

5 Konklúziók és további teendők

A cikkben egy párhuzamos korpuszon alapuló korpuszvezérelt megközelítést ismertettünk, amelyet kétnyelvű szótárak készítésének automatikus támogatására használtunk. A javasolt automatikus módszer lexikográfiai célokra számos ok miatt hasznosnak bizonyult. Ezek közül a legfontosabb, hogy – ha egy megfelelő méretű és reprezentatív korpusz rendelkezésre áll – a javasolt megközelítés garantálja, hogy a legrelevánsabb fordítások fognak szerepelni a szótárban. Ezért a javasolt módszer jobban kezeli a poliszémiát, mint akár a hagyományos lexikográfia, akár az egynyelvű korpuszokat felhasználó lexikográfia. Ezenfelül lehetővé válik a fordításjelöltek nyelvhasználaton alapuló rangsorolása: a legvalószínűbb fordításjelöltek szerepelnek először. A megközelítés további előnye, hogy az összes releváns példa könnyedén hozzáférhető, így a poliszém jelentések közül nagy mennyiségű természetes adat alapján választhatjuk ki a legmegfelelőbbet. A fenti tulajdonságok együttese különösen alkalmassá teszi az általunk javasolt módszert aktív szótárak előállítására.

Végül, a javasolt módszerrel könnyen előállíthatjuk a fordított irányú protoszótárat, hiszen csak a szóillesztő algoritmust kell újra alkalmazni.

A módszer hátrányai közé tartozik, hogy a kevéssé használt nyelvekre a megfelelő lefedettséget biztosító korpusz létrehozása rendkívül időigényes. Egyik fő feladatunk a litván-magyar párhuzamos korpusz méretének növelése.

Egy – a szóillesztő algoritmusból fakadó – további nehézség, hogy a módszer jelenlegi formájában nem alkalmas a többszavas kifejezések kezelésére. Egy lehetsé-

[13] A Bojtár-féle szótár valójában két fordítást ad meg, ám a második ezek közül csak a példamondatból derül ki.

[14] Hat javasolt fordításból négy volt jó.

ges megoldás a fordításjelöltekhez tartozó kontextusok alapján a megfelelő fordításokat az utószerkesztési munkálatok során kézzel hozzáadni. Egy további kutatási irányt képez a többszavas kifejezések automatikus kezelése.

Hivatkozások

1. Atkins, B. T. S., Rundell, M.: The Oxford Guide to Practical Lexicography. Oxford University Press (2008)
2. Bojtár E.: Litván-magyar nagyszótár. Akadémiai Kiadó, Budapest (2007)
3. Digitális Irodalmi Akadémia: http://www.pim.hu/
4. Erjavec, T., Ignat, C., Pouliquen, B., Steinberger, R.: Massive multi-lingual corpus compilation: Acquis Communautaire and totale. In: Proceedings of the 2nd Language & Technology Conference, April 21-23, 2005, Poznan, Poland (2005) 32-36
5. Magyar Elektronikus Könyvtár: http://mek.oszk.hu/
6. Och, F. J., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, Vol. 29, No. 1. (March 2003) 19-51
7. Oravecz, Cs., Dienes, P.: 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
8. Rimkutė, E., Daudaravičius, V., Utka, A., Kovalevskaitė, J.: Bilingual Parallel Corpora for English, Czech and Lithuanian. In: The Third Baltic Conference on Human Language Technologies 2007 Conference Proceedings. Kaunas (2008) 319–326
9. Rimkutė, E., Daudaravičius, V., A. Utka.: Morphological Annotation of the Lithuanian Corpus. In: 45th Annual Meeting of the Association for Computational Linguistics; Workshop Balto-Slavonic Natural Language Processing 2007 Conference Proceedings. Praga (2007) 94–99
10. Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V.: Parallel corpora for medium density languages. In: Proceedings of the RANLP 2005. (2005)