CONSTRUCȚIA AUTOMATĂ DE CORPUSURI MULTILINGUALE

TIBERIU BOROȘ, DAN TUFIȘ, ALEXANDRU CEAUȘU

Institutul pentru Cercetarea Inteligenței Aritificiale – Academia Română

{tibi,tufis,aceausu}@racai.ro

Rezumat

Insuficiența resurselor lingvistice pentru multe dintre limbile naturale este principalul impediment în progresul tehnologic al prelucrării automate a acestor limbi. Succesul abordărilor statistice în traducerea automată pentru limbile de largă circulație, bine echipate sub raport cantitativ și calitativ cu resurse lingvistice, a evidențiat încă o dată necesitatea colectării și preprocesării corespunzătoare a unor volume cât mai mari posibile de resurse lingvistice. Pentru aplicațiile multi- și cros-linguale corpusurile paralele și (mai recent) cele comparabile sunt resurse primare indispensabile. Articolul de față prezintă o serie de unelte ce pot fi folosite în extragerea automată de corpusuri paralele sau puternic comparabile.

1. Introducere

În ultimii 15-20 de ani abordările bazate pe tehnici inductive și volume mari de date, ca și creșterea spectaculoasă a performanțelor ce calcul și de stocare ale noilor generații de calculatoare au condus la progrese greu de anticipat în anii de început ai lingvisticii computaționale și ai traducerii automate. Utilizarea web-ului ca resursă primară de date, fără nici un fel de prelucrare lingvistică majoră, a demonstrat că utilizarea tehnologiilor statistice aplicate unor volume foarte mari de texte poate fi un răspuns la marea provocare a traducerii automate între toate limbile documentelor din spațiul web. Google translate[1] oferă deja traducere automată pentru 52 de limbi (2652 de perechi limbă sursă->limbă țintă). Sistemul Bing Translator[2] de la Microsoft oferă traducere automată pentru 30 de limbi (870 de perechi limbă sursă->limbă țintă). Pe lângă aceste exemple de sisteme publice foarte cunoscute pot fi amintite și companii mai noi[3], specializate pe traducere automată care dezvoltă sisteme comerciale focalizate pe domenii specializate, asigurând însă pentru unele perechi de limbi o calitate a traducerii aproape de nivelul traducerii umane[4].

Dacă metodele de prelucrare numerică bazate exclusiv pe texte neprelucrate lingvistic (în engleză acest tip de prelucrare este numit *number crunching NC*) sunt aplicabile pentru orice pereche de limbi, în schimb volumul datelor necesare pentru o calitate acceptabilă a traducerii este uriaș (pentru unele perechi de limbi nici măcar conținutul actual al web-ului nu este suficient), necesitând resurse de calcul și stocare ce depășesc dotările grupurilor obișnuite de cercetare. Numai marile companii își pot permite dezvoltări de sisteme de prelucrare multi- sau cros-lingvală pentru un număr mare de perechi arbitrare de limbi. Pe de altă parte cercetări recente (Koehn et al., 2007), (Hoang

[1] http://translate.google.com/# (consultare la data de 10 aprilie 2010)
[2] http://www.microsofttranslator.com/Default.aspx (consultare la data de 10 aprilie 2010)
[3] http://www.translationguide.com/translation_company_links.php (o listă parțială la data de 10 aprilie 2010)
[4] http://www.languageweaver.com/page/home/

et al.,2009), (Tufiş et al., 2009) ş.a. au arătat că atunci când textele sunt prelucrate lingvistic (segmentate, dezambiguizate morpho-lexical, lematizate, parsate) cu un volum de date mult mai mic decât conținutul (pentru o anumită pereche de limbi) întregului web se pot obține traduceri comparabile şi chiar superioare metodelor de tip NC. Fără îndoială însă că şi pentru astfel de abordări cu cât volumul datelor primare este mai mare calitatea traducerilor se îmbunătățeşte substanțial.

Pentru sistemele de traducere automată bazate pe metode statistice (indiferent dacă sunt de tip NC sau includ prelucrări lingvistice) cele mai valoroase resurse textuale sunt corpusurile paralele ce conțin texte în două sau mai multe limbi astfel încât ele reprezintă traduceri reciproce. Unul dintre primele corpusuri paralele bilingv (engleză-franceză) este corpusul Hansards conținând transcrieri ale dezbaterilor în Parlamentul Canadei în perioada 1975-1988 (Germann, 2001). El a constituit întotdeauna o resursă fundamentală pentru studiile cros-linguale pentru perechea de limbi engleză-franceză.

Multilingualismul este o caracteristică esențială a Europei şi implicațiile culturale, sociale şi economice au ridicat multilingualitatea la nivelul unei preocupări politice de prim rang şi absolut toate instituțiile Uniunii Europene produc un volum foarte mare de documente paralele. Necesitatea accesului cercetătorilor din domeniul prelucrării limbajelor naturale la aceste documente a determinat factorii responsabili (de exemplu OPOCE – biroul pentru publicații oficiale ale Uniunii Europene) să facă publice formatul electronic al unui volum din ce în ce mai mare de documente paralele pe baza cărora cercetătorii au construit primele corpusuri paralele europene publice. Primul dintre acestea a fost EuroParl (Koehn, 2005), disponibil pentru 11 din limbile comunității europene încă din 2001. El conține transcrieri ale sesiunilor din Parlamentul European din perioada 1996-2001. Versiunile următoare[5] au extins cantitatea de texte, dar nu şi numărul de limbi. Câțiva ani mai târziu, în 2006, a fost creat şi distribuit prima variantă a corpusului JRC-Acquis (Steinberger et al., 2006). Versiunea V3 este în prezent[6] cel mai mare corpus multilingual disponibil, conținând texte de natură juridică în 22 de limbi ale Uniunii Europene. Acest corpus conține peste 1 miliard de cuvinte, în medie 48 milioane de cuvinte/limbă. Din perspectiva limbii române relevante alături de JRC-Acquis mai sunt relevante corpusurile EMEA şi OpenSubs (Tiedemann, 2009). EMEA conține documente ale Agenție Europene a Medicamentului în 22 de limbi, iar OpenSubs conține subtitrări în 30 de limbi. Textele româneşti din JRC-Acquis şi EMEA conțin diacritice în schimb ele lipsesc din OpenSubs ceea ce face acest ultim corpus mai puțin util pentru prelucrarea limbii române.

La o privire generală, cele mai multe corpusuri paralele au două mari probleme inerente: conțin, în principal, limbi de largă circulație şi sunt specifice unui anumit domeniu.

Folosirea de corpusuri comparabile devine o alternativă viabilă pentru antrenarea sistemelor de traducere în cazul domeniilor şi limbilor care au o mică reprezentare în corpusuri paralele. Un corpus comparabil reprezintă o colecție de texte în două sau mai multe limbi care deşi nu reprezintă traduceri reciproce riguroase, conțin informații similare. Aceste tipuri de corpusuri au grade de comparabilitate diferite (Skandina et al.,

[5] Versiunea v5, distribuită la începutul anului 2010, acoperă perioada 1996-2009 şi conține aproximativ 55 milioane de cuvinte pentru fiecare din cele 11 limbi: franceză, italiană, spaniolă, portugheză, engleză, olandeză, germană, daneză, suedeză, gracă şi finlandeză.

[6] http://langtech.jrc.it/JRC-Acquis.html

2010), cele mai utile fiind cele clasificate drept „puternic comparabile" (eng. *strongly comparable*). Aceste corpusuri comparabile conţin informaţii despre aceleaşi lucruri, folosesc acelaşi registru lingvistic şi au un grad ridicat de suprapunere la nivelul echivalenţilor lexicali de traducere. Deşi antrenarea sistemelor de traducere folosind corpusuri comparabile implică un nivel de complexitate ridicat faţă de antrenarea tradiţională ce foloseşte resurse paralele, textele comparabile sunt disponibile în proporţie mult mai mare decât cele paralele.

2. *Combinarea modelelor de traducere extrase din corpusuri paralele şi corpusuri comparabile*

În cadrul proiectului naţional STAR (PNII – IDEI 742/19.01.2009) la Institutul de Cercetări pentru Inteligenţa Artificială a fost realizat prototipul unui sistem de traducere pentru perechea de limbi engleză-română (Ceauşu, 2009) folosind platforma Moses (Koehn et al, 2007) şi serviciile proprii de procesare a textului (Tufiş et al, 2007). Sistemul a fost antrenat pe un corpus de aproape un milion de unităţi de traducere cu peste treizeci de milioane de atomi lexicali. Corpusul este compus în proporţie de 75% din texte din domeniul juridic, 5% texte din domeniul jurnalistic. Restul de 20% sunt echivalenţi de traducere şi unităţi de traducere extrase din ontologia lexicală românească Ro-Wordnet extinsă cu terminologie juridică (Tufiş et al, 2008). Experimentele noastre au arătat că traducerea textelor din domeniul juridic are o calitate bună dar, deşi inteligibile, traducerile textelor din alte domenii (de ex. sport, medicină, turism etc.) au o calitate mult mai slabă. O soluţie raţională pentru remedierea acestui aspect constă în construcţia mai multor modele statistice de limbă şi de traducere, fiecare caracteristic unui anumit registru textual. În continuare, un modul preliminar ar putea clasifica un text nou ce urmează a fi tradus într-una din clasele pentru care există modele de traducere şi sistemul va efectua traducerea folosind modelul specific. O altă variantă constă în a combina diferitele modelele, atribuind dinamic, în funcţie de domeniul textului ce urmează a fi tradus, ponderi de influenţă diferite modelelor statistice combinate.

Proiectul european Accurat[7] (248347/FP7), lansat la începutul acestui an, are ca obiectiv, printre altele, construcţia unor corpusuri comparabile cât mai mari pentru care să poată complementa puţinele corpusuri paralele disponibile pentru limbile proiectului (inclusiv româna). Combinarea modelelor de traducere construite din cele două categorii de corpusuri se va face studiind variantele amintite anterior şi alegând soluţia cea mai performantă.

Pentru a extinde sistemul de traducere şi pentru alt domeniu decât cel juridic, decodorului Moses îi pot fi adăugate modele de traducere antrenate folosind corpusurile comparabile.

Moses este un decodor pentru sistemele de traducere automată care extinde traducerea formei de ocurenţă a cuvintelor cu modele de traducere factorizate. Spre deosebire de decodarea bazată pe echivalenţii de traducere constituiţi din secvenţe contigue de cuvinte, care se bazează numai pe forma de ocurenţă a cuvintelor, traducerea factorizată

[7] http://www.accurat-project.eu/

poate lua în considerare informații suplimentare asociate atomului lexical, cum ar fi partea de vorbire, forma de dicționar a cuvântului sau descrierea sa morfo-sintactică.

În conformitate cu modelul traducerii factorizate, procesul de traducere presupune căutarea traducerii *t* (în limba țintă) care maximizează o combinație liniară a probabilităților diverşilor factori utilizați. Probabilitatea de traducere şi regula de decizie este dată de:

$$t^* = \arg\max_{e} \sum_{k=1}^{n} \lambda_k h_k(t, s) \quad (1)$$

unde h_k (t,s) este unul din cei *n* factori (o funcție de trăsături caracteristice perechii <t,s>) iar λ_k este ponderea acestuia. Cei mai importanți factori care contribuie la probabilitatea de traducere provin din modelele de traducere, modelele de limbă, modelele de generare şi cele de distorsiune. Combinarea modelelor de traducere din corpusuri comparabile şi a modelelor de traducere din corpusuri paralele este dată de ponderile λ_k folosite pentru fiecare factor.

Stabilirea ponderilor λ_k se realizează prin proceduri de învățare automată. Una dintre cele mai utilizate proceduri este MERT (Minimal Error Rate Training) (Och, 2003) care presupune existența unor traduceri de referință, în raport cu care se induc valorile λ_k pentru care diferența dintre textul de tradus şi textul de referință este minimă în raport cu o măsură de similaritate. Algoritmul MERT este un proces iterativ, computațional foarte intens, dar complet nesupervizat. Decodorul MOSES este însoțit în distribuția standard şi de programul care implementează algoritmul MERT care implicit utilizează scorul BLEU (Papineni, et al., 2002). Această implementare (Bertoldi et al, 2009) a algoritmului MERT poate folosi opțional şi alte măsuri de evaluare.

Pentru colectarea corpusurilor necesare proiectelor noastre a fost construit un program specializat care va fi descris în continuare. După cum se va vedea din prezentarea sistemului, acesta poate colecta atât corpusuri paralele cât şi corpusuri comparabile, în funcție de natura conținutului multilingual al site-urilor prelucrate.

3. Descrierea sistemului de colectare a corpusurilor multilingve

O componentă importantă a motoarelor de căutare este un modul (numit *spider* sau *crawler*) care citeşte fiecare pagină şi reține legăturile care pleacă din aceasta (Kobayashi and Takeda, 2000). Astfel se generează un graf plecând de la o listă inițială de adrese ce conțin referințe către alte locații. Procesul de generare poate fi automat (prin liste de căutare generate la prima indexare urmând ca la anumite intervale de timp să se verifice schimbări în conținutul acestora), manual (lista de site-uri este actualizată prin intervenție umană) sau hibrid.

Problema care apare în cazul extragerii de corpus este relevanța rezultatelor. În multe situații se obțin şi referințe către site-uri fără nici o legătură cu tema de la care s-a plecat şi pentru care nu se mai pot aplica aceleaşi reguli de indexare.

Din acest motiv, listele trebuie supuse unui proces de filtrare în funcție de specificul fiecărei pagini. Stabilirea link-urilor ce vor fi verificate se face prin filtrarea acestora conform unor criterii bine alese. De exemplu:

– Dacă se doreşte ca articolul/documentul vizitat să facă parte din pagina curentă se poate recurge la eliminarea elementelor ***href*** ce conțin ***http://*** în interior (majoritatea referințelor interne se fac prin calea relativă către document) sau eventuala eliminare a legăturilor ce nu conțin adresa de bază a site-ului curent.
– Există cazuri în care se poate face o selectare bazată pe organizarea internă a paginii web. De exemplu, toate paginile de ştiri conțin în adresele lor ***/news/*** iar toate paginile referitoare la vreme conțin în adresele lor ***/weather/***

Indiferent de metoda care se alege, dacă structura sitului se modifică la un moment dat, aplicația trebuie actualizată corespunzător, atrăgând după sine modificări în codul sursă, care devine greu de citit şi reutilizat în alte aplicații.

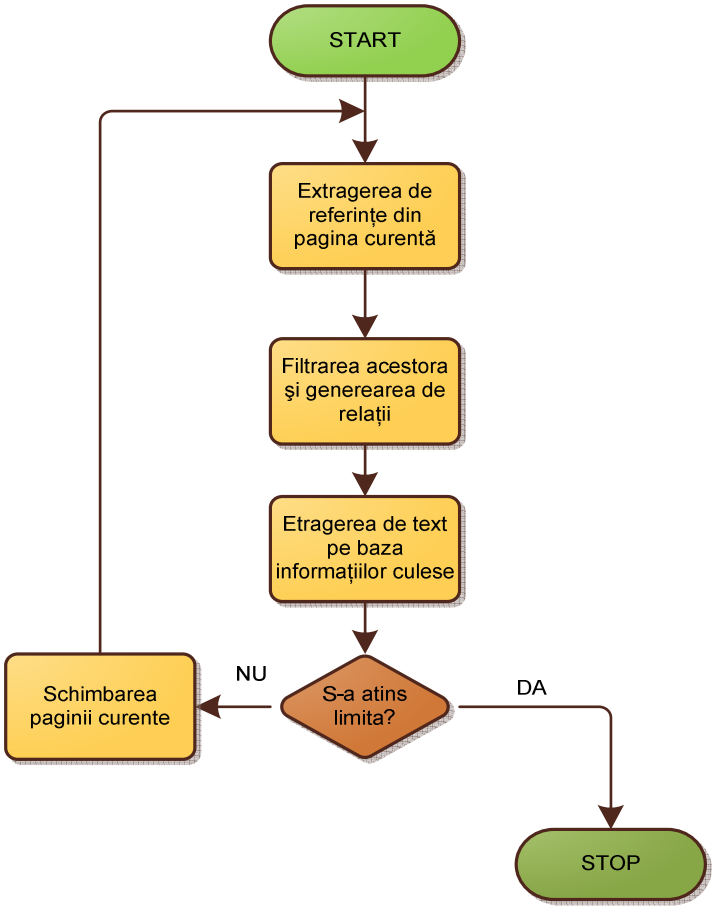

Figura 1: Diagrama de flux pentru algoritmul de extragere de text din pagini web

În continuare propunem o soluție de simplificare şi minimizare a efortului depus la adăugarea unor noi criterii de indexare sau modificarea celor deja existente, prin utilizarea unei arhitecturi bazate pe diagrame de flux. Fiecare bloc al diagramei de flux este independent de celelalte putând fi editat în orice moment şi permițându-se adăugarea de paşi suplimentari la fluxul deja existent. În fundal (eng. *background)* sunt lansate procese de tip consolă care primesc ca parametrii de intrare fluxul şi generează

ieşire de tip text. Aceasta este preluată şi este transmisă, în urma aplicării unei alte serii de prelucrări, la modulul următor.

În varianta actuală a sistemului am implementat două tipuri de blocuri, ilustrate în Figura 1, acestea fiind suficiente pentru operațiile uzuale într-un flux de date:

- blocuri de procesare – module ce vor genera sau altera textul
- blocuri decizionale – stabilesc drumul care va fi urmat

3.1. Arhitectura sistemului

Sistemul de colectare de corpusuri multilinguale este compus din două module distincte: editorul de fluxuri şi serviciul de colectare a paginilor web.

EDITORUL DE FLUXURI. În cadrul editorului de fluxuri se realizează schema de funcționare a aplicației, iar editarea diagramei poate fi făcută în mod vizual. De asemenea, pot fi editați parametrii lansării în execuție a proceselor ce vor alcătui sistemul de extragere, iar blocurile pot fi testate în mod individual.

SERVICIUL DE COLECTARE A PAGINILOR WEB este un serviciu Windows care rulează la un interval de timp prestabilit. Parametrii de configurare a serviciului sunt încărcați dintr-un fişier XML. Documentul XML reprezintă schema de funcționare a aplicației creată cu ajutorul editorului de fluxuri. Serviciul de colectare a paginilor web, conform diagramei, lansează în execuție procesele şi asigură comunicarea între ele. De asemenea, serviciul dispune de un jurnal detaliat al operațiilor efectuate cu ajutorul căruia pot fi investigate posibilele erori ale proceselor lansate.

3.2. Modul de utilizare

Pentru colectarea de pe web a unui corpus multilingual într-un anumit domeniu este necesară crearea unei diagrame cu ajutorul aplicației de editare de fluxuri. În această diagramă sunt specificate procesele ce urmează a fi executate precum şi condițiile în care acestea pot fi executate. De asemenea, în diagramă trebuie stabiliți parametrii de lansare în execuție a proceselor (interpretorul utilizat, parametrii în linie de comandă etc.). Condiționările se specifică prin expresii regulate vor fi aplicate intrării/ieşirii proceselor. Pentru algoritmii în care este luată o decizie, în diagramă se va completa şi regula de validare a conținutului. De pildă, după descărcarea unui document el poate fi clasificat automat şi, în raport cu categoria identificată, el poate fi stocat într-un director conținând documente similare sau eventual şters în cazul în care categoria sa nu este de interes sau nu este identificată cu suficientă precizie.

Fiecare bloc din diagramă reprezintă o unitate de execuție. Aceasta este caracterizată de calea către numele programului ce urmează a fi lansat, parametrii liniei de comandă (cuvintele cheie $script şi $input vor fi înlocuite cu numele script-ului şi, respectiv, intrarea obținută la pasul anterior) şi expresiile regulate folosite (pentru intrare, ieşire şi condiția de validare).

Odată definită, diagrama de flux este salvată într-un fişier XML ce este importat de sistemul propriu-zis de extragere. Acesta execută întregul flux la intervalul de timp

stabilit în prealabil. Diagrama poate fi modificată în orice moment, urmând ca aceasta să fie reîncărcată la următoarea pornire a aplicației de colectare de pagini web.

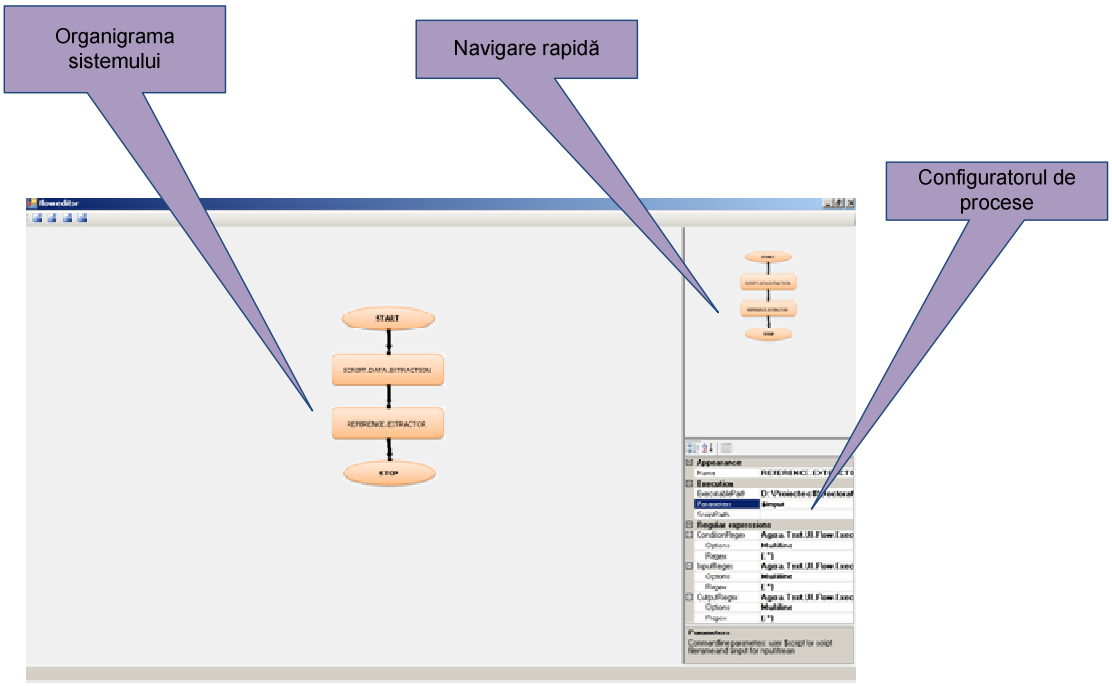

Figura 2: Editorul de diagrame

4. *Cazuri practice*

Prezentăm în continuare două exemple de utilizare a aplicației: (i) pentru constituirea de corpus multilingual din domeniul turismului (*Wikitravel*) şi pentru colectarea de corpus puternic comparabil din domeniul jurnalistic (*Parlamentul European* - secțiunea de ştiri). În ambele cazuri s-au folosit scheme de procesare simple similare cu cea prezentată în Figura 2. Au fost utilizate doar două programe/procese şi nu au fost necesare blocuri decizionale. Cele două procese constau în: (i) generarea unei liste de adrese ce urmează a fi scanate şi (ii) parcurgerea acestora şi prelucrarea rezultatelor.

4.1. Wikitravel

În cazul site-ului Wikitravel s-a urmărit extragerea de corpusuri puternic comparabile din registrul textual al turismului pentru limbile română, engleză şi germană. Cu ajutorul unei expresii regulate au fost extrase adresele paginilor din limba română disponibile în domeniul „wikitravel.org”. Au fost urmate link-urile ce se regăseau în versiunea pentru limba română a paginii, urmând ca traducerile să fie găsite prin căutarea textului în subdomeniile ***/de/*** sau ***/en/***. Pentru validarea adreselor candidate au fost folosite liste pentru secvențele de adresă permise şi pentru secvențele de adresă nepermise, prioritatea pentru validare fiind acordată secvențelor permise. Legăturile care reprezentau traduceri ale articolului pentru engleză şi germană au fost grupate, iar în cazul în care acestea lipseau, au fost înlocuite generic cu o pagină goală. Procesul a fost continuat prin extragerea textului şi eliminarea marcajelor HTML. În Tabelul 2 sunt prezentate cantitativ rezultatele extragerii de texte în limbile de interes. Valorile din

tabel reprezintă numărul aproximativ de cuvinte pentru fiecare limbă din perechea respectivă.

Tabel 1: Reguli de validare a adreselor

Secvențe de adresă permise	Secvențe de adresă nepermise
wikitravel.org/de/	#
wikitravel.org/en/	mailto
wikitravel.org/ro/	special
	action=
	Special:
	doubleclick
	User:
	Wikitravel:
	Utilizator:
	Image:
	http:
	www.

Tabel 2: Rezultatele obținute în cazul Wikitravel

română - engleză	română-germană
>400.000 cuvinte	>100.000 cuvinte

Dinamica site-ului wikitravel.org nu este foarte dinamică astfel că după o creştere rapidă a volumului datelor descărcate, extensia corpusului a devenit foarte lentă.

4.2. Parlamentul European

Pe site-ul Parlamentului European la secțiunea ştiri se găsesc articole traduse în 22 de limbi. La fel ca în cazul Wikitravel, pornind de la adresa *http://www.europarl.europa.eu/news/public/toute_actualite/default/default_ro.htm* se parcurg toate referințele găsite pe pagină. Spre deosebire de cazul anterior, un site de ştiri se actualizează mai des, astfel că datele cantitative prezentate în Tabelul 3 sunt cele de la momentul scrierii acestui articol.

Tabel 3: Corpus obținut

Limbă	Cod	Cuvinte
bulgară	bg	54351
cehă	cs	57480
daneză	da	67856
germană	de	70307
greacă	el	90729
engleză	en	84484
spaniolă	es	91447
estoniană	et	39388
finlandeză	fi	62692
franceză	fr	89171
ungară	hu	59584

Limbă	Cod	Cuvinte
italiană	it	80431
lituaniană	lt	80431
letonă	lv	67211
malteză	mt	46041
olandeză	nl	74390
poloneză	pl	56031
portugheză	pt	75557
română	ro	81433
slovacă	sk	67189
slovenă	sv	68794

Sperăm ca în câteva luni, corpusul multilingual EU-News (22 limbi) să devină o resursă publică, de dimensiune corespunzătoare, pentru cercetări şi dezvoltări în domeniul cros- şi multi-lingual (traducere automată, sisteme de regăsire documentară multilingvală, sisteme de întrebare/răspuns cros-linguale, etc.). În ambele exemple prezentate, corpusurile au fost stocate în directoare distincte de fişiere de tip text. Documentele paralele/comparabile au fost numite folosind identificatori unici extensiile acestora denotând codul ISO al limbii documentului (de pildă nume de tipul xxxxxxxxx.bg, xxxxxxxxx.cs, ..., xxxxxxxxx.sv identifică fişierele conţinînd documente paralele/comparabile în cele 22 de limbi ale corpusului multilingual). În plus, au fost generate metadate corespunzătoare fiecărui fisier text, metadate ce documentează calea către fişier, adresa de la care a fost extras fişierul, numărul de cuvinte din fişier etc.

5. Concluzii

Rezultatele bune obţinute în urma colectării de corpus multilingual din site-ul Wikitravel şi din secţiunea de ştiri a site-ului Parlamentului European dovedesc utilitatea setului de instrumente propus în această lucrare. Aceste instrumente facilitează interoperabilitatea diverselor aplicaţii scrise în diferite limbaje de programare, fiecare aplicaţie fiind reprezentată ca un proces al unei diagrame de flux. De asemenea, posibilitatea de a reconfigura în orice moment o diagramă de flux permite abordarea unor proiecte de colectare de corpus de complexitate ridicată – procesele, şi regulile de validare pot fi schimbate chiar în cursul colectării documentelor.

Mulţumiri. Activitatea de cercetare descrisă a fost sprijinită de Comisia Europeană prin proiectul ACCURAT (248347/FP7) şi (parţial) de CNCSIS – UEFISCSU prin proiectul PNII – IDEI STAR, 742/19.01.2009

Referinţe bibliografice

Bertoldi, N., Haddow, B., Fouet, J.-B. (2009): *Improved Minimum Error Rate Training in Moses*, în Prague Bulletin of Mathematical Linguistics, Nr. 91, 2009, pp. 7-16

Ceauşu, A. (2009) *Tehnici de traducere automată şi aplicabilitatea lor limbii române ca limbă sursă*, Bucureşti, România: Teză de doctorat, Institutul de Cercetări pentru Inteligenţă Artificială, Academia Română.

Germann, U. (ed.) (2001). *Aligned Hansards of the 36th Parliament of Canada* - Release 2001-1a. http://www.isi.edu/natural-language/download/hansard/

Hoang, H., Koehn, P., Lopez, A. (2009: A Uniform Framework for Phrase-Based, Hierarchical and Syntax-Based Machine Translation, International Workshop on Machine Translation (IWSLT), pp. 152-159

Koehn, Ph. (2005). *EuroParl: A Parallel Corpus for Statistical Machine Translation*. Machine Translation Summit 2005. Phuket, Thailand. http://people.csail.mit.edu/koehn/publications/europarl/

Koehn, Ph., Hoang, H., Birch, A., Callison-Burch, Ch., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, Ch., Zens, R., Dyer, Ch., Bojar, O., Constantin, A., Herbst, E. (2007). *Moses: Open Source Toolkit for Statistical Machine*

Translation, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007

Kobayashi, M. and Takeda, K. (2000). "Information retrieval on the web". *ACM Computing Surveys* (ACM Press) **32** (2): 144–173. doi:10.1145/358923.358934

Och, F. J. (2003). *Minimal Error Rate Training in Statistical Machine Translation*. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, July 2003, pp. 160-167

Papineni, K., Roukos, S., Ward, T., and Zhu, W. J. (2002): *BLEU: a method for automatic evaluation of machine translation*. In ACL-2002: 40th Annual meeting of the Association for Computational Linguistics pp. 311–318

Skadina, I., Vasiljevs, A., Skadinš, R., , Gaizauskas, R., Tufiş, D., Gornostay, T. (2010): Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation. Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, May, Malta.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006): *The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages.* In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pp. 2142-2147, Genoa, Italy, May 2006. ELRA - European Language Ressources Association.

Tiedemann, J. (2009), *News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces*. In N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) *Recent Advances in Natural Language Processing* (vol V), pages 237-248, John Benjamins, Amsterdam/Philadelphia

Tufiş, D., Ion, R., Bozianu, L., Ceauşu, A., and Ştefănescu, D. (2008): *Romanian Wordnet: Current State, New Applications and Prospects.* In Attila Tanacs, Dora Csendes, Veronika Vincze, Christiane Fellbaum, and Piek Vossen (eds.), *Proceedings of 4th Global WordNet Conference, GWC-2008*, pp. 441-452, Szeged, Hungary, January 2008. University of Szeged, Hungary. ISBN 978-963-482-854-9

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2007): *Servicii web lingvistice ale ICIA*. In Ionuț Pistol, Dan Cristea, and Tufiş, D. (eds.), Resurse Lingvistice şi Instrumente pentru Prelucrarea Limbii Române, pp. 61-68, Iaşi, România, dec. 2007. Universitatea "Al.I. Cuza" Iaşi, Editura Universității "Al.I. Cuza" Iaşi. ISBN 978-97-3703-297-3

Tufiş, D., Koeva, S., Erjavec, T., Gavrilidou, M., Krstev, C. (2009): Building Language Resources and Traslation Models for Machine Translation Focused on South Slavic and Balkan Languages. In J. Machacova and K. Rohsmann (eds)"Scientific Results of the SEE-ERA.NET Pilot Joint Call". Center for Social Innovation Publisher, Vienna, ISBN 978-3-200-01567-8, pp. 37-48, June 2009