Radu Ion, Alexandru Ceauşu, Dan Ştefănescu, and Dan Tufiş. Servicii web interoperabile şi multilinguale. In Adrian Iftene, Horia-Nicolai Teodorescu, Dan Cristea, and Dan Tufiş (eds.) Resurse Lingvistice şi Instrumente pentru Prelucrarea Limbii Române, pp. 175-184, Iaşi, Romania, septembrie 2010. Universitatea "Al.I. Cuza" Iaşi, Editura Universităţii "Al.I. Cuza" Iaşi. ISSN 1843-911X.

SERVICII WEB INTEROPERABILE ŞI MULTILINGUALE

RADU ION, ALEXANDRU CEAUŞU, DAN ŞTEFĂNESCU, DAN TUFIŞ

Institutul de Cercetări pentru Inteligență Artificială, Academia Română

{radu, aceausu, danstef, tufis}@racai.ro

Rezumat

Problemele interoperabilității instrumentelor şi resurselor lingvistice sunt preocupări majore ale cercetării actuale în domeniul prelucrării limbajelor naturale. Cele mai noi tehnologii, bazate pe arhitecturi orientate pe servicii web, constituie un pas important în direcția asigurării interoperabilității. Atunci când serviciile web sunt independente de limbă sau facil adaptabile la diverse limbi criteriile de interoperabilitate şi multilingualitate sunt în mare măsură satisfăcute. Orchestrarea diverselor servicii se face de obicei prin alegerea unui format comun al datelor de intrare/ieşire. În acest sens, au apărut deja o serie de platforme de integrare a resurselor şi instrumentelor lingvistice dezvoltate şi rezidente în locații diferite, implementate în limbaje de programare şi sub sisteme de operare variate. În această lucrare prezentăm câteva dintre serviciile web ale Institutului de Cercetări pentru Inteligență Artificială (ICIA) adaptate la platforma de servicii web WebLicht. WebLicht reprezintă un mediu în care serviciile web înregistrate pot interacționa unele cu altele datorită conversiei parametrilor de intrare/ieşire la formatul TCF (engl. Text Corpus Format).

1. Introducere

Tehnologiile Web Service, un palier fundamental în filosofia noilor generații ale web-ului, înlesnesc utilizatorilor dezvoltarea de aplicații ce integrează diverse module, implementate de autori diferiți, în limbaje diferite, şi chiar aflate pe alte maşini decât cea locală. Conceptul de arhitectură orientată spre servicii (SOA) a apărut dintr-o nevoie imperioasă de a structura şi standardiza colecția eterogenă şi vastă de instrumente şi documente care există în spațiul virtual al intra- şi inter-rețelelor informatice locale, regionale sau mondiale. Derivat din conceptul SOA, dar mai general, este conceptul de infrastructură de servicii web care oferă răspunsuri la construcția de fluxuri de prelucrare pe baza serviciilor web dar şi la identificarea serviciilor relevante pentru o anumită aplicație, a datelor necesare diferitelor servicii, a publicării de resurse şi instrumente de prelucrare a acestor resurse, a documentării lor etc. O astfel de infrastructură este în curs de construcție în cadrul proiectului european CLARIN[1].

CLARIN îşi propune să reunească aplicații şi resurse din domeniul Prelucrării Automate a Limbajului Natural (PLN) şi să le prezinte sub o formă în care nespecialiştii în prelucrarea limbajului natural (PLN) să le poată identifica şi utiliza cât mai uşor în activitățile pe care le întreprind. Altfel spus, CLARIN urmăreşte realizarea unei infrastructuri europene a tehnologiilor limbajului natural (scris şi vorbit) astfel încât

[1] http://www.clarin.eu

acestea să fie *integrate* (aplicațiile şi resursele se află în centre specializate care utilizează servere dedicate interconectate prin rețele de tip GRID), *interoperabile* (resursele şi serviciile vor fi descrise cu limbajele proprii Web-ului semantic (engl. Semantic Web) pentru a depăşi barierele puse de diversitatea formatelor de intrare/ieşire), *stabile* (garantarea funcționării neîntrerupte şi existența personalului de suport tehnic), *persistente* (garantarea funcționării continue pe o perioadă lungă de timp – cel puțin pe durata proiectului), *accesibile* (resursele şi aplicațiile sunt accesibile online pe Web) şi *extensibile* (întreg mediul trebuie să incorporeze uşor noi aplicații şi resurse). CLARIN reuneşte şi armonizează la nivel european o multitudine de proiecte naționale. Astfel de proiecte sunt, printre altele, proiectul german D-SPIN şi proiectul românesc CLARIN-RO.

D-SPIN[2] (Hinrichs et al., 2008) a implementat o platformă de servicii web care să asigure interoperabilitatea unor aplicații de PLN dezvoltate independent. Metoda de interconectare a fost aceea a expunerii acestor aplicații ca servicii web de tip REST[3] (şi nu SOAP[4] pentru că s-a considerat că volumul de lucru suplimentar necesar împachetării/despachetării formatului SOAP va introduce un timp suplimentar şi inutil de adaptare pentru noile aplicații care doresc înregistrarea). Această platformă se numeşte **WebLicht**[5] şi a fost dezvoltată în colaborare de universitățile din Tübingen, Stuttgart, Leipzig şi Berlin. În prezent conține peste 70 de aplicații de PLN (pentru limbile germană, engleză, franceză, română, spaniolă, italiană şi finlandeză) expuse ca servicii web şi care *sunt compatibile între ele ca formate de intrare/ieşire*. Acest lucru permite în mod evident posibilitatea extraordinară a compunerii de operații simple pentru obținerea unor rezultate care altfel ar fi fost imposibil de obținut – de exemplu, combinarea adnotării cu etichete morfosintactice a unui text din două surse (programe) diferite. Dintre operațiile ce se pot efectua asupra unui text putem menționa: segmentare lexicală, adnotare cu etichete morfo-sintactice, lematizare, analiză sintactică etc.

Formatul parametrilor de intrare/ieşire este unul XML care a fost definit special pentru a facilita integrarea aplicațiilor care prelucrează texte. Acesta se numeşte Text Corpus Format (TCF, (Schmid, 2009)) şi reprezintă o stivă de adnotări care sunt produse de operații elementare de procesare a textelor (de exemplu, segmentarea la nivel de unitate lexicală este o adnotare necesară operației de etichetare morfo-sintactică). Astfel, fiecare serviciu web din WebLicht aşteaptă la intrare un fişier XML care conține textul prelucrat până la un anumit nivel şi întoarce acelaşi fişier XML la care adaugă un nou nivel care conține adnotările pe care le produce (niciunei operații nu-i este permis să şteargă vreun nivel de adnotare). Menționăm că platforma Weblicht este una dintre componentele ECD (European Clarin Demonstrator) prototipul funcțional al etapei de specificare (preparatory phase) a proiectului CLARIN, prototip a cărui implementare este coordonată de ICIA.

[2] http://weblicht.sfs.uni-tuebingen.de/englisch/index.shtml
[3] Representational State Transfer, un tip de comunicare de tip client-server.
[4] Ajuns la versiunea 1.2, SOAP (http://www.w3.org/TR/soap12-part0/) este un protocol descris în XML cu care diverse aplicații distribuite pe calculatoare conectate în rețea, pot comunica prin schimb de mesaje împachetate în formatul SOAP.
[5] http://weblicht.sfs.uni-tuebingen.de:8080/WebLicht1/

În cele ce urmează, vom descrie procesul de adaptare, în cadrul proiectului RO-CLARIN, a câtorva dintre serviciile web ale ICIA la platforma WebLicht. O serie de alte servicii web proprii sau în curs de implementare la ICIA vor fi adăugate în viitorul apropiat platformei WebLicht. De asemenea vom exemplifica avantaje ale adoptării de standarde pentru uşurinţa adaptării unor aplicaţii diverse (expuse ca servicii web) la rezolvarea unor probleme care altfel nu ar fi putut fi abordate. În acest sens, vom arăta cum anume putem combina rezultatele mai multor programe de etichetare morfo-sintactică operând pe aceeaşi segmentare lexicală cu metoda descrisă în (Tufiş, 1999) care opera cu acelaşi program de etichetare morfo-sintactică antrenat însă pe diverse corpusuri.

2. TTL în platforma WebLicht

TTL[6] este un modul Perl care oferă următoarele adnotări ale textelor în limbile română, engleză şi franceză: segmentare la nivel de frază şi unitate lexicală, adnotare cu etichete morfo-sintactice (engl. POS tagging), lematizare, recunoaşterea entităţilor denumite (engl. named entity recognition) şi analiză sintactică de suprafaţă (engl. chunking). A ajuns la versiunea 7.9 şi a fost implementat deja ca serviciu web (Tufiş et al., 2007).

Pentru adaptarea TTL la WebLicht a fost nevoie de următoarele modificări asupra implementării anterioare ca serviciu web:

- renunţarea la suportul pentru SOAP întrucât WebLicht specifică comunicarea de tip REST ca fiind standardul acceptat. În acest caz, un client care doreşte să folosească serviciul web va trimite o cerere HTTP POST către serverul de web care găzduieşte serviciul comunicându-i acestuia un şir de caractere codificat UTF-8 care reprezintă un document XML TCF bine format. Serviciul web procesează documentul şi îl returnează clientului cu stratul de adnotări suplimentar pe care l-a produs.
- renunţarea la serverul de web implementat de pachetul Perl **SOAP::Lite** şi adoptarea serverului de web **Apache**[7] ca gazdă pentru TTL. În acest fel am rezolvat o problemă gravă a serverului de web din SOAP::Lite care nu accepta decât o singură conexiune la serviciul web la un moment dat şi care de altfel, era şi instabil. Cu Apache a trebuit să utilizăm modulul Fast CGI[8] care permite serverului de web să încarce *o singură dată* o instanţă a serviciului web (moment în care se încarcă toate resursele necesare – operaţie costisitoare ca timp de execuţie) iar apoi să servească mai mulţi clienţi folosind această instanţă. De asemenea, FCGI permite distribuţia uniformă a încărcării pe instanţele serviciului web existente şi un întreg management al pornirii/opririi instanţelor serviciului web.

Serviciul web TTL pentru platforma WebLicht expune câte un URL pentru fiecare operaţie elementară pentru fiecare limbă în parte (`{lang}` poate lua una din valorile engleză (`en`), română (`ro`) sau franceză (`fr`)):

[6] Abrevierea din engleză pentru "Tokenizing, Tagging and Lemmatizing free running texts".
[7] http://httpd.apache.org/
[8] FCGI, http://www.fastcgi.com/

- segmentare la nivel de frază şi apoi la nivel lexical: `http://ws2.racai.ro/TTL-{lang}-tokenizer`;
- adnotare cu etichete morfo-sintactice: `http://ws2.racai.ro/TTL-{lang}-postagger`;
- lematizare: `http://ws2.racai.ro/TTL-{lang}-lemmatizer`;
- analiza sintactică de suprafață: `http://ws2.racai.ro/TTL-{lang}-chunker`;

În afară de aceste operații, orice serviciu web din WebLicht necesită conversia textelor primare (care nu sunt adnotate în vreun fel) la formatul TCF de bază (fişierul XML care este preluat de prima operație care, de obicei, este segmentarea la nivel de frază/unitate lexicală). În prezent, WebLicht conține astfel de servicii web care convertesc texte UTF-8 şi RTF la formatul TCF de bază exemplificat în figura 1.

```
<?xml version="1.0" encoding="UTF-8"?>
<D-Spin xmlns="http://www.dspin.de/data" version="0.3">
  <tns:MetaData xmlns:tns="http://www.dspin.de/data/metadata">
    <tns:source></tns:source>
  </tns:MetaData>
  <tns:TextCorpus xmlns:tns="http://www.dspin.de/data/textcorpus" lang="ro">
    <tns:text>Aceasta este o propoziție de test.</tns:text>
  </tns:TextCorpus>
</D-Spin>
```

Figura 1: Formatul TCF pentru propoziția „*Aceasta este o propoziție de test.*”

Figura 2 reprezintă o sesiune de lucru în WebLicht. Primul pas pe care trebuie să-l facă utilizatorul este să selecteze limba textului pentru care dorește procesarea. Apoi, alegând orice operație din coloana din partea stângă, sistemul va adăuga automat în lanțul de prelucrare toate operațiile care aduc textul din forma brută în formatul TCF necesar la intrarea operației selectate de utilizator. De exemplu, dacă am fi selectat operația de adnotare cu etichete morfo-sintactice, WebLicht ne-ar fi adăugat automat în lanțul de prelucrare operațiile de conversie la formatul TCF de bază (cel din figura 1) şi segmentare la nivel de frază şi unitate lexicală.

Lanțul de prelucrare final apare în chenarul „Selected Tools:” iar apăsarea butonului „Run” produce rezultatele vizibile în partea de jos a coloanei din partea dreaptă. Acestea se pot descărca de pe site în formatul XML TCF produs de lanțul de prelucrare.

Anterior aminteam de posibilitatea extraordinară de a combina diverse operații WebLicht pentru a obține rezultate care altfel (în absența acestei platforme) ar fi fost destul de dificil (sau imposibil în cazul în care aplicațiile nu erau disponibile) de obținut. Studiul de caz este combinarea rezultatelor a două adnotatoare cu etichete morfo-sintactice care rulează pe acelaşi text segmentat lexical şi frazal în acelaşi fel pentru ambele programe. Tufiş (1999) discută posibilitatea obținerii unei adnotări morfo-sintactice mai bune în cazul în care acelaşi program de adnotare morfo-sintactică rulează cu modele de limbă diferite. WebLicht ne permite însă să obținem o adnotare mai bună (aplicând deci aceleaşi procedee ca în lucrarea citată – calculul matricelor de confuzie) folosind două adnotatoare diferite: în cazul nostru, TTL şi TreeTagger pentru limba engleză. În figura 3 avem rezultatele celor două programe pentru propoziția „*A very simple test sentence is the test bed for the combined classifiers model.*”

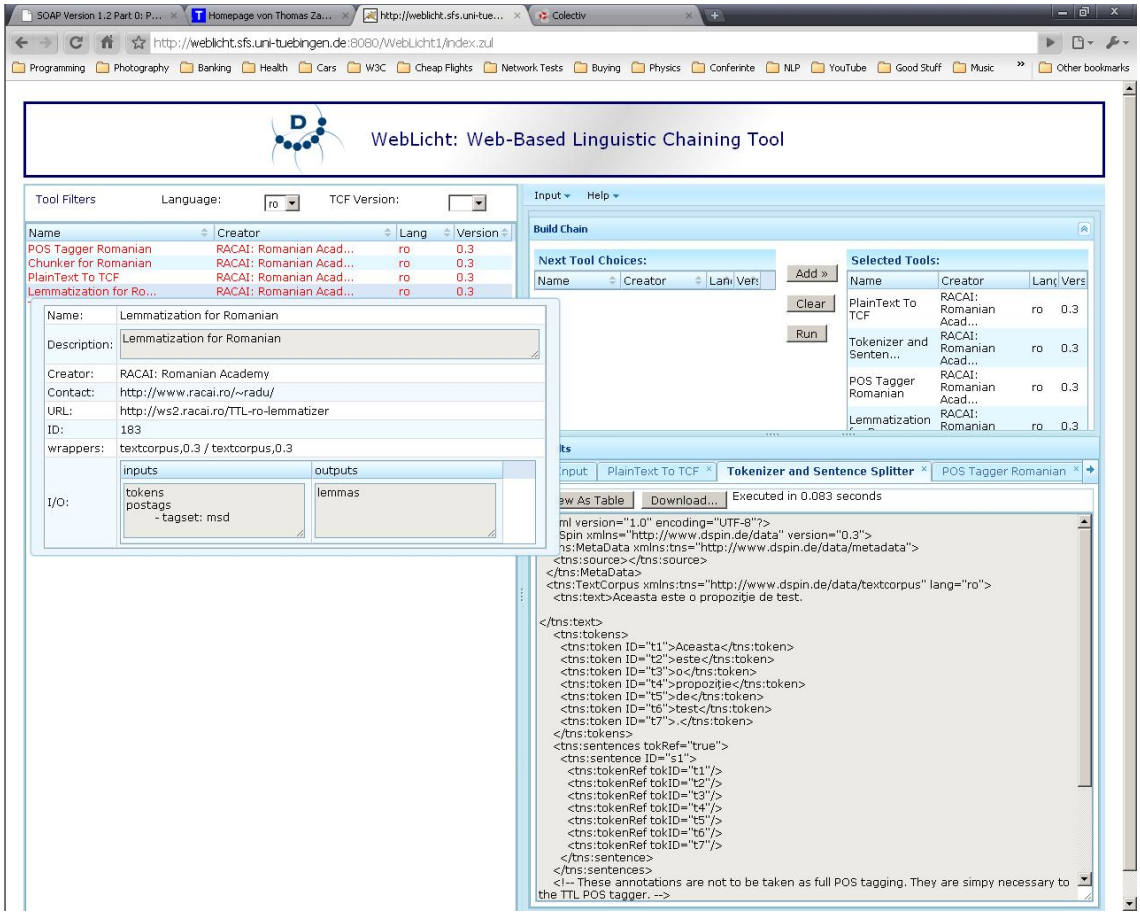

Figura 2: O sesiune de lucru în WebLicht

Pentru a reuşi combinarea rezultatelor avem nevoie de corespondența etichetelor folosite (o altă aplicație extrem de utilă pe care WebLicht o face posibilă). TTL foloseşte mulțimea de etichete Multext-East (MSD) iar TreeTagger-ul foloseşte mulțimea de etichete Penn Treebank. Ştim de exemplu că substantivele la singular se codifică cu „NN” în Penn Treebank şi cu „Ncns” în MSD, substantivele la plural cu „NNS” respectiv „Ncnp”, adjectivele cu „JJ” respectiv „Afp”, ș.a.m.d. În exemplul considerat, ambele programe au găsit adnotarea corectă, dar, în cazul textelor mari, acest lucru este puțin probabil iar cu tehnica din (Tufiş, 1999) (după ce am fixat corespondența etichetelor), putem obține rezultate mai bune.

Tokenization	TreeTagger	TTL
A	**DT**	**Ti-s**
very	**RB**	**Rsp**
simple	**JJ**	**Afp**
test	**NN**	**Ncns**
sentence	**NN**	**Ncns**
is	**VBZ**	**Vmip3s**
the	**DT**	**Dd**
test	**NN**	**Ncns**
bed	**NN**	**Ncns**
for	**IN**	**Sp**
the	**DT**	**Dd**
combined	**JJ**	**Afp**
classifiers	**NNS**	**Ncnp**
model	**NN**	**Ncns**
.	.	**PERIOD**

Figura 3: Rezultatele adnotatoarelor TTL şi TreeTagger pe o propoziție de test

3. *Servicii REST de procesare a textului bazate pe metode statistice de maximizare a entropiei*

Pe lângă TTL, ICIA mai dispune de aplicații performante de prelucrare primară a textelor care au fost de asemenea adaptate ca servicii web SOAP aşa cum se raportează în (Tufiş et al., 2007). Ca o noutate față de aceste servicii web de tip SOAP dezvoltate folosind platforma .Net Framework[9] pentru engleză şi română, noile servicii conforme cu specificațiile WebLicht (apel de tip REST şi parametri de intrare/ieşire în format TCF) sunt disponibile şi pentru limbile franceză şi germană. Adaptarea la noile limbi a fost posibilă datorită flexibilității metodei de antrenare statistică, serviciul REST de adnotare morfo-sintactică[10] folosind modulul de adnotare stratificată METT – Maximum Entropy Tiered Tagging (Ceauşu, 2006), pe principiul maximizării entropiei, similar modelului ME al lui Ratnaparkhi (1988). Ca şi TTL, tagger-ul METT utilizează pentru fiecare limbă un set de descrieri morfo-sintactice compatibil cu specificațiile Multex-East şi un tagset redus pentru adnotarea stratificată (Tufiş & Dragomirescu, 2004).

[9] http://www.microsoft.com/NET/

[10] http://www.racai.ro/RestWS/Service.svc/ws-{lang}-postagger unde „{lang}” poate lua valorile „en”, „ro”, „fr” şi „de”.

Noile servicii web de procesare a textului care implementează METT pentru WebLicht sunt disponibile la adresa:

`http://www.racai.ro/RestWS/Service.svc/ws-{lang}-{webservice}`

unde `{lang}` este limba textului ce urmează a fi procesat, iar `{webservice}` este tipul de procesare ce urmează a fi invocat ca serviciu web. `{lang}` poate lua valorie română (`ro`), engleză (`en`), franceză (`fr`) şi germană (`de`) iar `{webservice}` poate fi înlocuit cu unul din următoarele tipuri de prelucrări: segmentare lexicală (`tokenizer`), segmentare la nivel de frază (`sentsplitter`), adnotare morfo-sintactică (`postagger`) şi lematizare (`lemmatizer`). Trebuie să notăm faptul că este obligatoriu ca în orice interogare HTTP la unul din aceste URL-uri, să specificăm în preambulul interogării tipul "application/xml" pentru fişierul TCF pe care îl vom trimite spre prelucrare (în antetul „Content-type").

Pentru conversia în formatul TCF folosim serviciul web:

`http://www.racai.ro/RestWS/Service.svc/converter-{lang}`

unde `{lang}` este codul ISO de două caractere al limbii textului care urmează a fi procesat. Dacă limba nu este precizată, va fi invocat automat serviciul web pentru recunoaşterea limbii (descris în secțiunea 4).

În figura 4 este prezentat un exemplu de interogare a serviciilor web pentru conversia de la text la formatul TCF, segmentare lexicală şi segmentare la nivel de frază, adnotare morfo-sintactică şi lematizare. În continuare, folosind exemplul din figura 4, descriem parametrii de intrare şi ieşire necesare fiecărui serviciu.

La interogarea serviciului web de conversie a unui text în formatul TCF (http://www.racai.ro/RestWS/Service.svc/converter-ro) rezultatul va fi un document XML similar celui din figura 1. Noul document XML va avea sub elementul „tns:TextCorpus" doar un element „tns:text" cu textul conținut în interogarea HTTP. În cazul acestui serviciu, parametrul de intrare este de tipul „text/plain" (specificat în „Content-type") iar cel de ieşire este de tipul „application/xml".

Pentru următoarea etapă de procesare – segmentarea lexicală – rezultatul serviciului de conversie este folosit ca argument pentru serviciul web http://www.racai.ro/RestWS/Service.svc/ws-ro-tokenizer. Rezultatul interogării HTTP este un document XML care are, pe lângă elementul „tns:text", şi un alt element „tns:tokens" care conține atomii lexicali din „tns:text". Fiecare element „tns:token" conține indexul din textul inițial şi lungimea atomului lexical.

În continuare, documentul XML rezultat al serviciului de segmentare lexicală este folosit ca parametru de intrare pentru serviciul de segmentare la nivel de frază (http://www.racai.ro/RestWS/Service.svc/ws-ro-sentsplitter). Acest serviciu adaugă elementul „tns:sentences" în care sunt grupate elemente „tns:sentence" reprezentând frazele formate din atomii lexicali din „tns:tokens". Serviciul de segmentare frazală necesită prezența unui element „tns:tokens" în documentul XML inițial, deoarece fiecare frază conține referințe către atomii lexicali.

În final, rezultatul serviciului de segmentare la nivel de frază este folosit ca parametru de intrare pentru serviciul de adnotare morfo-sintactică stratificată

(http://www.racai.ro/RestWS/Service.svc/ws-ro-postagger). Deşi poate folosi ca parametru de intrare un document XML conținând doar segmentarea lexicală (cu element „tns:tokens”), este recomandat ca documentul XML ce urmează a fi prelucrat să conțină şi segmentarea frazală (cu element „tns:sentences”), informația suplimentară contribuind la o viteză crescută a adnotării. Acest serviciu adaugă documentului XML inițial un nou element „tns:POStags” în care sunt precizate descrierile morfo-sintactice ale atomilor lexicali.

Serviciul web de lematizare (http://www.racai.ro/RestWS/Service.svc/ws-ro-lematizer) necesită rezultatul tokenizării şi al adnotării morfo-sintactice pentru a atribui o lemă fiecărui atom lexical. Acest serviciu adaugă un element „tns:lemmas” documentului XML inițial.

```
<?xml version="1.0" encoding="utf-8" ?>
- <D-Spin version="3.0" xmlns="http://www.dspin.de/data"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:xsd="http://www.w3.org/2001/XMLSchema">
    <tns:MetaData xmlns:tns="http://www.dspin.de/data/metadata" />
  - <tns:TextCorpus lang="ro"
      xmlns:tns="http://www.dspin.de/data/textcorpus">
      <tns:text>Acesta este un exemplu.</tns:text>
    - <tns:tokens>
        <tns:token ID="t1" start="0" end="6">Acesta</tns:token>
        <tns:token ID="t2" start="7" end="11">este</tns:token>
        <tns:token ID="t3" start="12" end="14">un</tns:token>
        <tns:token ID="t4" start="15" end="22">exemplu</tns:token>
        <tns:token ID="t5" start="22" end="23">.</tns:token>
      </tns:tokens>
    - <tns:sentences>
      - <tns:sentence ID="s1">
          <tns:tokenRef tokID="t1" />
          <tns:tokenRef tokID="t2" />
          <tns:tokenRef tokID="t3" />
          <tns:tokenRef tokID="t4" />
          <tns:tokenRef tokID="t5" />
        </tns:sentence>
      </tns:sentences>
    - <tns:POStags tagset="MSD">
        <tns:tag tokID="t1">Pd3msr</tns:tag>
        <tns:tag tokID="t2">Vmip3s</tns:tag>
        <tns:tag tokID="t3">Timsr</tns:tag>
        <tns:tag tokID="t4">Ncms-n</tns:tag>
        <tns:tag tokID="t5">PERIOD</tns:tag>
      </tns:POStags>
    - <tns:lemmas>
        <tns:lemma tokID="t1">acesta</tns:lemma>
        <tns:lemma tokID="t2">fi</tns:lemma>
        <tns:lemma tokID="t3">un</tns:lemma>
        <tns:lemma tokID="t4">exemplu</tns:lemma>
        <tns:lemma tokID="t5">.</tns:lemma>
      </tns:lemmas>
    </tns:TextCorpus>
</D-Spin>
```

Figura 4: Exemplu de rezultat al interogării serviciilor web de identificare a limbii, de segmentare lexicală şi frazală, de adnotare morfo-sintactică şi de lematizare.

4. Identificarea limbii

Acest serviciu[11] asigură identificarea automată a limbii unui text scris într-una dintre cele 22 de limbi ale Uniunii Europene. Textul ar trebui să conţină un număr minim de 10-15 cuvinte (în principiu, o propoziţie). Serviciul web de identificare a limbii funcţionează ca un convertor din text în formatul TCF care poate recunoaşte automat limba textului, adăugând un atribut „lang" elementului „tns:TextCorpus".

Identificarea limbii este făcută folosind modele statistice pentru fiecare limbă şi un modul de predicţie. Predicţia este realizată prin calcularea unui scor de similaritate al textului de intrare cu fiecare model de limbă în parte. Modelele de limbă se realizează pe baza ponderii pe care o au prefixele şi sufixele cuvintelor în textele de antrenare disponibile pentru limba respectivă.

În experimentele realizate până acum, am utilizat texte de antrenament (cu mărimi ce au variat între 0,5 şi 1,2 MB) pentru cele 22 limbi oficiale ale Acquis-ului Communautaire (Steinberger et al, 2006). Textele, fiind însă din domeniul juridic şi având o structură mai aparte[12], nu sunt tocmai reprezentative pentru limbile luate în discuţie. Cu toate acestea, am obţinut rezultate excelente folosind pentru prefixe o lungime de trei caractere iar pentru sufixe de patru.

5. Concluzii

De la intrarea lor în funcţiune pe data de 19 februarie 2010, serviciile web ale ICIA din cadrul platformei WebLicht au avut 1389 de accesări pentru toate operaţiile expuse pentru toate limbile. Dintre acestea cele mai multe accesări au fost de test (câte aprox. 40 de bytes de text pe cerere) dar au fost şi cazuri în care s-a cerut procesarea a mai mult de 2KB de text. Suntem astfel siguri că efortul de integrare a aplicaţiilor noastre în standardele promovate de CLARIN va fi de folos atât nouă cât şi comunităţii CLARIN care creşte într-un ritm alert.

Proiectul CLARIN deschide o nouă cale în abordarea cercetărilor în Prelucrarea Automată a Limbajului Natural şi Lingvistică Computaţională prin operaţiunile de standardizare, colectare şi diseminare a unor colecţii impresionante de resurse şi aplicaţii ale acestor domenii care altfel ar fi fost în marea majoritate a cazurilor, inaccesibile. Astfel, noi probleme de cercetare pot să apară sau unele mai vechi îşi pot găsi rezolvarea. În orice caz, adoptarea standardelor CLARIN va asigura accesul neîngrădit al celor neiniţiaţi la tehnologiile limbajului pe de o parte şi îmbunătăţirea considerabilă a şanselor cercetătorilor de a găsi rapid soluţii la problemele lor de cealaltă.

Mulţumiri. Activitatea de cercetare descrisă în această lucrare a fost sprijinită de proiectul european FP7 „CLARIN – Common Language Resources and Technology Infrastructure" (nr. de proiect 212230) finanţat de Comisia Europeană şi de proiectul

[11] `http://www.racai.ro/RestWS/Service.svc/converter`

[12] Textele juridice au adesea o structură formată din multe aliniate în care anumiţi termeni se repetă de foarte multe ori afectând acoperirea lingvistică a modelelor de limbă.

românesc PC7 „CLARIN–RO: Infrastructură pentru resurse lingvistice interoperabile pentru limba română” (nr. de proiect 16EU/06.04.2009) finanţat de ANCS.

Referinţe bibliografice

Ceauşu, A. (2006). Maximum Entropy Tiered Tagging. Janneke Huitink & Sophia Katrenko (editors), *Proceedings of the Eleventh ESSLLI Student Session, ESSLLI 2006*, June 2006, Malaga, Spain, pp. 173—179.

Hinrichs, E., Wittenburg, P., Lemnitzer, L., Geyken, A. (2008). D-SPIN - the German CLARIN initiative. In *CLARIN Newsletter #2*, 2008 (http://www.clarin.eu/files/cnl02-web.pdf).

Kemps-Snider, M., Bel, N., Broeder, D. (2009). Proposal for a CLARIN Service CMDI Components. September, 2009, http://www.clarin.eu/wp2/wg-26/wg-26-documents/cmdi-profile-for-web-services

Kemps-Snider, M., Bel, N. (2009). CLARIN Report on Web Services. March 2009 http://www.clarin.eu/wp2/wg-26

Ratnaparkhi, A. (1998). Maximum Entropy Models for Natural Language Ambiguity Resolution. Ph.D. thesis, University of Pennsylvania, Philadelphia.

Schmid, H. (2009). The technical details of the D-SPIN Architecture. Internal technical report, 2009 (http://weblicht.sfs.uni-tuebingen.de/englisch/publikationen.shtml).

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th LREC Conference*, Genoa, Italy, 22-28 May, 2006, pp.2142-2147, ISBN 2-9517408-2-4, EAN 9782951740822.

Tufiş, D. (1999). Tiered Tagging and Combined Language Models Classifiers. In Václav Matousek, Pavel Mautner, Jana Ocelíková, and Petr Sojka, editors, *Text, Speech and Dialogue (TSD 1999)*, Lecture Notes in Artificial Intelligence 1692, pp. 28—33. Springer Berlin / Heidelberg, January 1999. ISBN 978-3-540-66494-9.

Tufiş, D., Dragomirescu, L. (2004). Tiered Tagging Revisited. In *Proceedings of the 4th LREC Conference*, Lisabona, 2004, pp. 39—42.

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2007). Servicii Web lingvistice ale ICIA. În Ionuţ Cristian Pistol, Dan Cristea, şi Dan Tufiş, editori, *Lucrările atelierului RESURSE LINGVISTICE ŞI INSTRUMENTE PENTRU PRELUCRAREA LIMBII ROMÂNE*, pp. 61–68, Iaşi, România, 14–15 decembrie 2007. Editura Universităţii "Alexandru Ioan Cuza" Iaşi.