WS4LR - a Worksation for Lexical Resources

Cvetana Krstev, Ranka Stanković, Duško Vitas, Ivan Obradović

WS4LR: A Workstation for Lexical Resources

Cvetana Krstev[1], Ranka Stanković[2], Duško Vitas[3] and Ivan Obradović[2]

[1]Faculty of Philology, Studentski trg 3, CS-11000 Belgrade,

[2]Faculty of Mining and Geology, Đušina 7, CS-11000 Belgrade,

[3]Faculty of Mathematics, Studentski trg 16, CS-11000 Belgrade

E-mail: cvetana@matf.bg.ac.yu, ranka@rgf.bg.ac.yu, vitas@matf.bg.ac.yu, ivano@afrodita.rcub.bg.ac.yu

Abstract

In this paper we describe WS4LR, the workstation for lexical resources, a software tool developed within the Human Language Technology Group at the Faculty of Mathematics, University of Belgrade. The tool is aimed at manipulating heterogeneous lexical resources, and the need for such a tool came from the large volume of resources the Group has developed in the course of many years and within different projects. The tool handles morphological dictionaries, wordnets, aligned texts and transducers equally and has already proved very useful for various tasks. Although it has so far been used mainly for Serbian, WS4LR is not language dependent and can be successfully used for resources in other languages provided that they follow the described formats and methodologies. The tool operates on the .NET platform and runs on a personal computer under Windows 2000/XP/2003 operating system with at least 256MB of internal memory.

1 Introduction

The Human Language Technology group at the Faculty of Mathematics has been developing various lexical resources over quite a long period, reaching a considerable volume to date. Given the fact that these resources have been developed for many years, they have naturally been conceived within different projects and frameworks, both from the conceptual and the technological point of view. Although the HLT group made every reasonable effort to keep the ever growing pool of resources as coherent and standardized as possible, a certain level of heterogeneity was inevitable. Hence, due to the growth of the volume of resources as well as their heterogeneity, there was a rising need for developing a tool that would facilitate the maintenance, exploitation and integration of available resources as well as their further development. Embarking on this task, the HLT group has recently produced an integrated and easily adjustable tool, a workstation for language resources, labeled WS4LR, which greatly enhances the potentials of manipulating each particular resource as well as several resources simultaneously.

The paper is organized as follows: in section 2 we describe the lexical resources that can be handled by WS4LR, in section 3 we present the WS4LR modules and their functions, in section 4 some software consideration are given, and section 5 offers some conclusions and ideas for further work.

2. Overview of Resources

Various lexical resources that can be produced and handled by WS4LR are briefly described in this section.

2.1 Morphological Dictionaries

Morphological dictionaries of simple words and compounds[1] in LADL format (Courtois & Silberztein, 1990) exist for many languages, including French, English, Greek, Portuguese, Russian, Thai, Korean, Italian, Spanish, Norwegian, Arabic, German, Polish, Bulgarian, and Serbian. The Intex[2] , Unitex[3] and Nooj[4] systems for natural language processing based on linguistic resources provide for text processing using this type of dictionaries, but offer no facilities for dictionary development and management. WS4LR enables manipulation of dictionaries both of lemmas and of inflected forms.

In LADL format, all the entries in the dictionary of simple word lemmas, the so called DELAS, have the following form:

lemma.Knnn [+*SinSem*]*

where *lemma* is the simple word, in general in the form usually used in traditional dictionaries, *K* is the part of speech mark, *nnn* is the number denoting the class of lemmas that all share the same inflectional properties described by the appropriate transducer *Knnn*, and *+SinSem* is the freely attached marker that describes the syntactic, semantic, derivational, or other properties of a lemma. A part of speech code and an inflectional class code uniquely determine the finite transducer that generates all the forms in a lemma paradigm. A finite transducer, being capable of producing the output, adds to all these forms their possible grammatical categories. The DELAS dictionary and the set of transducers describing inflectional properties are used to produce the morphological dictionary of word forms, the so called DELAF. All the entries in this dictionary have the following form:

form,lemma[*:categories*]*

where *form* is a simple word form of a *lemma* that is represented by its DELAS entry form, and *:categories* are

[1] Term multi-word unit is sometimes used.
[2] Intex homepage: http://msh.univ-fcomte.fr/intex/
[3] Unitex homepage: http://www-igm.univ-mlv.fr/~unitex/
[4] Nooj homepage: http://www.nooj4nlp.net

the possible grammatical categories of the word *form*, each category represented by a single character code.

Morphological dictionaries of compound lemmas and word forms, named DELAC and DELACF, follow a similar format, except that both *lemma* and *form* can contain non-alphabetic characters – blank, hyphen, apostrophe, and alike. There is, however, one substantial difference. The inflection of compounds is more complex: in order to obtain all forms of a compound two different types of information are necessary. The first one deals with the inflection of simple lemmas that constitute a compound, while the other governs how these inflected forms combine in order to obtain the inflected form of a compound by taking into consideration grammatical agreement, word order, omissions, etc. The entry in DELAC is therefore accordingly more complex:

c-lemma.Cnnn [*+SinSem*]*

where *c-lemma* is a list of constituents in the form of entries of the DELAF dictionary of simple word forms:

c-lemma=lemma$_1$*.Nnnn*[*:categories*],
lemma$_2$*.Nnnn*[*:categories*],
lemma$_3$*.Nnnn*[*:categories*],...

The code *Cnnn* identifies a new type of transducer responsible for the inflection of compounds, described in (Savary, 2005). Examples of entries in these dictionaries are given in Appendix A.

2.2 Wordnets

Roughly speaking, a wordnet, such as the Princeton *WordNet* (PWN) is composed of synsets, or sets of synonymous words representing a concept, with basic semantic relations between them forming a semantic network (Fellbaum, 1998). Each synset word or "literal" is denoted by a "literal string" followed by a "sense tag" which represents the specific sense of the literal string in that synset, pretty much as in any explanatory dictionary, where an entry corresponding to a word is followed by number of its possible meanings. Following the basic principles set by PWN, wordnets for many other languages were developed, some in the scope of international projects, others independently. An important impact to the wordnet development came from the EuroWordnet project (Vossen, 1998) where the idea of an Interlingual index (ILI) has been introduced that enables the connection of the same concepts in different languages. The usage of ILI was further explored in the scope of the Balkanet project (Stamou, 2002) where all wordnets were synchronously developed on the basis of PWN. Within this project an implicit wordnet XML scheme was developed and further used in a number of software tools for wordnet management, such as VidDic (Horák, 2004). The underlying XML format is illustrated in Appendix B.

2.3 Aligned Texts

A pair of *semantically* equivalent texts in different langauges, such as an original text and its translation, that are and *aligned* on a structural level (paragraph, sentence, phrase, etc.) is known as an *aligned text* or *bitext*. Aligned texts are usually constructed in two main steps: in the first step, the texts to be aligned are segmented into equivalent units, and in the second step the correspondence between these units is established. The equivalent units are usually sentences, but the units can be larger, as well as smaller.

The standard method for representing aligned texts is the Translation Memory eXchange format (TMX) that is XML-compliant [5]. The alignment itself can be performed by different methods and tools (Veronis, 2000). Of particular interest are programs that use XML tagged input texts and produce the result also as an XML document. Such is the case with XAlign[6]. In Appendix C short examples of input and output to XAlign, as well as the corresponding TMX format are given.

2.4 Finite Transducers

As described in section 2.1, inflectional paradigms are represented by appropriate *finite state transducers* usually produced by the graph management tool in Intex/Unitex environment. The produced graphs are in the form of standard textual files that can easily be generated or managed independently of these systems and their respective graph management tools. The same is true for other types of graphs produced and used within Intex/Unitex.

2.5 Bilingual Lists

As a result of a various translation and lexicographic projects various unstructured *bilingual wordlists* from various domains were produced. An excerpt from one such list is presented in Appendix E.

3. WS4LR Organization

WS4LR is organized in modules which perform different functions as depicted in Figure 1.

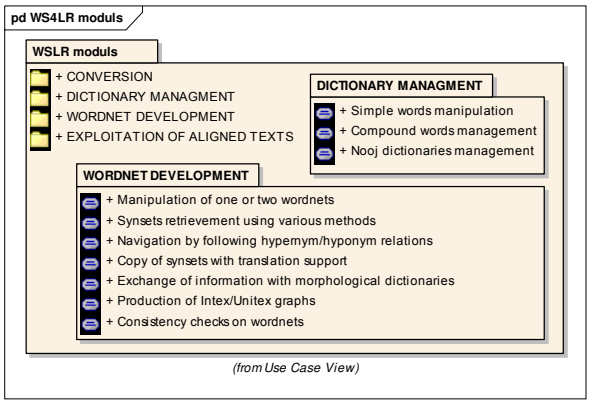

Figure 1. The UML diagram with WS4LR modules

3.1 Conversion

This module enables the user to perform a conversion from one character encoding set to another on a chosen set of files (e.g. all files in one directory). The user can choose a conversion Perl or awk script suitable for the specific file type, or produce his/her own script easily.

This module makes switching between Intex and Unitex easy. This would otherwise be a problem since Intex does not support Unicode and Unitext works only with Unicode. This feature is particularly useful for

[5] TMX 1.4b Specification, OSCAR Recommendation, 26 April 2005, http://www.lisa.org/tmx/

[6] Language and Dialog page of Loria: http://led.loria.fr/collab.php#projet45

Serbian where two alphabets, Cyrillic and Latin, are used, and lexical and textual resources must exist for both. To that end the HLT group produces resources for Serbian in a special encoding that uses the ASCII character set and that can be unambiguously transformed into Serbian Latin or Serbian Cyrillic alphabet. As noted before, WS4LR offers to the user the option to apply the transformation only to a part of the file, such as an XML file where only the text should be converted while the XML tags shouldn't be altered. Similarly, when a DELAS or DELAF type file is transformed, only lemmas and word forms are converted, not the part of speech and grammatical codes.

3.2 Dictionary Management

This module enables concurrent manipulation of a set of dictionaries of lemmas, simple words or compounds, distributed in several files. Working with DELAF type files is not directly supported since this type of files should in general be produced automatically from DELAS by applying the appropriate transducers. The organization of dictionaries in separate files is important from the practical point of view since smaller files are easier to manipulate. An even more important reason is the fact that in text recognition by Intex/Unitex the usage of all dictionaries is not always necessary, or even recommended. For example, dictionaries of English personal names transcribed according to Serbian orthography should not be applied to a text that makes no reference to such persons, since that could only unnecessarily add to text ambiguity.

An important feature of this module is the ability of retrieving efficiently a subset of lemmas by matching the lemmas, their Part-of-Speech, inflectional class code, syntactic and semantic markers or their Boolean combination. For instance, one can look for all the dictionary entries starting or ending with a search string. The later option illustrated in Figure 2 is particularly useful when the inflectional code of a new lemma is being established, since it depends on its ending.

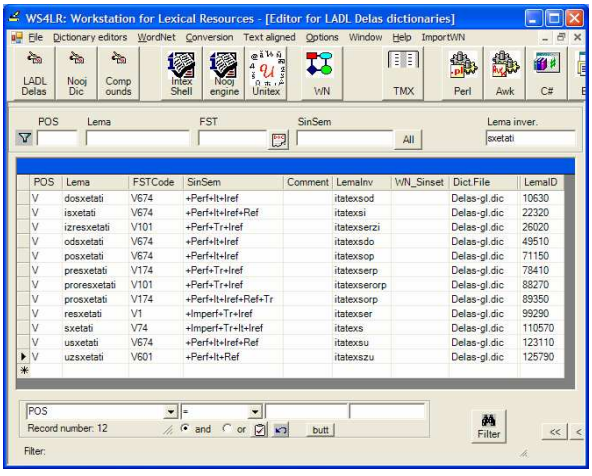

Figure 2. Retrieval of all verbs ending with *sxetati*: they all belong to two main inflectional groups: V1 and V74.

This module enables the user to modify or delete all the information attached to a lemma, or the lemma itself, as well as to add new entries. A new entry can be generated from scratch or by copying an existing lemma, which in some cases facilitates the work. The regular expression or FSA graph describing the inflectional properties of the selected lemma can be inspected and corrected if found inadequate.

The handling of dictionaries of compound lemmas is similar, though with some important differences. For instance, the search by the entry ending is not supported, since it does not make sense in this case. The form for new entries is more complex since more information need to be supplied. In the upper part of the form the information pertaining to the entry as a whole is displayed or typed, while in the lower part the information associated to the compound lemma constituents is entered (Figure 3). For inflected compound constituents additional information is needed: the lemma, its inflection class, as well as the list of grammatical categories of the form that appears in the compound lemma. For example, in the compound *turska kafa* (Engl. Turkish coffee), the lemma for the constituent form *turska* is *turski.* The form of this adjective in the compound lemma is inflected in order to agree in gender with the noun *kafa*.

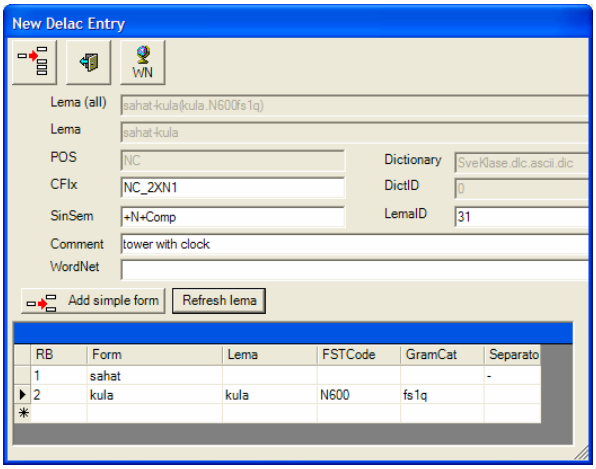

Figure 3. The form for compound entries

3.3 Wordnet Manipulation

This module supports the manipulation of wordnets. The user can work with one wordnet only, or with two wordnets, which can then be synchronized using ILI. The user can navigate through wordnets by following hypernym/hyponym relations. Synsets can be retrieved using various methods, from simple string matching to complex Xpath expressions, either predefined or specified by the user. For instance, by means of the Xpath expression "//SYNSET[DOMAIN='geology']" the user can retrieve all synsets from the working wordnet that belong to the domain of geology, or more precisely, that contain the element <DOMAIN> with the content "geology". New synsets can be added to wordnets using predefined forms. If working simultaneously with two wordnets, the user can copy a synset form one wordnet to another thus synchronizing them automatically via the ILI. Unstructured bilingual lists may be used to suggest possible candidates for a synset. The module also performs various consistency checks on wordnets such as detecting dangling relations.

A particularly interesting feature of this tool is that it enables the exchange of information between wordnets and morphological dictionaries. Namely, morphosyntactic information from dictionaries can be attached to synset literals. The tool searches for the wordnet literal in dictionaries of simple or compound lemmas, and it retrieves from them its inflectional class code. If more lemmas of the same form exist, they are all offered to the user to choose the appropriate one. Conversely, semantic marks of synset literals can be assigned to dictionary entries (Krstev & al., 2004). For instance, the mark +Comm can be added to all communicative verbs, that is, all literals belonging to the synsets that are hyponyms of the synset <communicate:2, intercommunicate:2> can obtain this mark in the morphological dictionary. The module enables easy production of Intex/Unitex graphs that locate all literals from a chosen synset in a text, with or without synset hypernyms.

3.4 Working with Aligned Texts

The module uses texts which have previously been aligned using Xalign as an alignment tool and converts them to TMX format, or texts that are already in that format. By choosing the appropriate XSLT stylesheet various visualizations can be obtained, in HTML or other formats.

Powerful linguistic tools such as Intex/Unitex, though inherently multilingual since resources for them have been developed for many languages, presently do not support simultaneous work with different languages. With WS4LR we have tried to, at least partially, overcome this shortcoming and enable better exploitation of aligned texts as resources of great value. This is achieved by an integration of all the resources supported by WS4LR. It may best be illustrated by concordance production using various search criteria such as simple strings, lemmas (with all their inflectional forms) or concepts (all or some literals from chosen synsets and/or their hypernyms) (Figure 4). In aligned segments retrieved, occurrences that correspond to search criteria in the source language are highlighted (Figure 5).

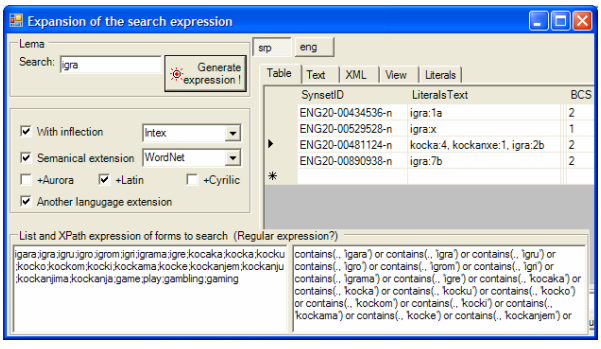

Figure 4. The form for expansion of the search criteria

The user can also use the translation equivalence option which is aimed at locating equivalences in target language for occurrences found in the source language. This is done on the basis of data from wordnets for corresponding languages. This option can be particularly useful in further development of wordnets, since aligned segments where a translation equivalent could not be located potentially contain a concept not yet covered by the corresponding wordnet.

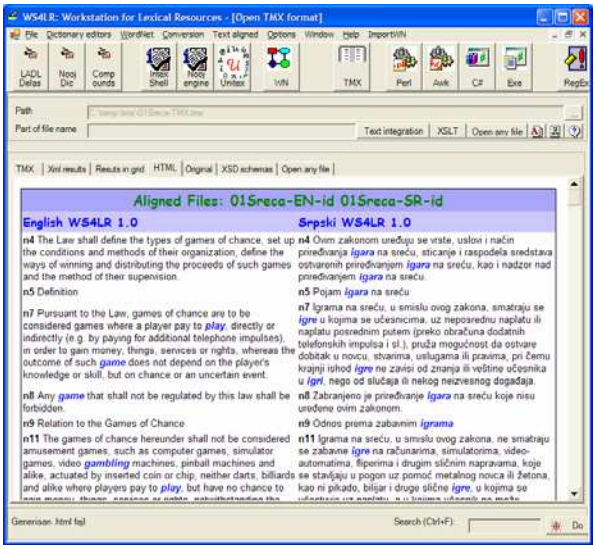

Figure 5. Retrieved aligned segments with highlighted occurrences that correspond to search criteria

4. Programming Considerations

WS4LR is written in C#, operates on the .NET platform and can run on any personal computer under Windows 2000/XP/2003 operating system with at least 256MB (preferably 512MB) of internal memory. The solution consists of five projects, the main exe ConvertCP project and four .Net libraries of classes, as depicted in Figure 6.

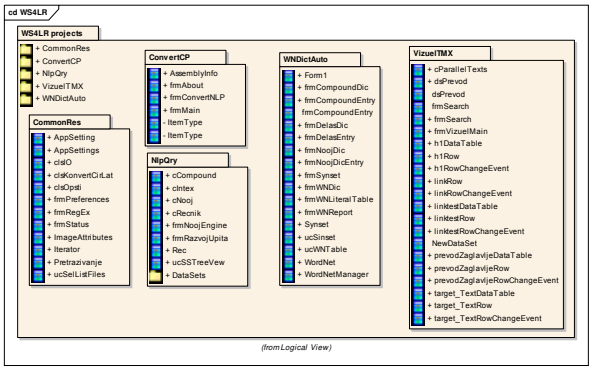

Figure 6. Project components of WS4LR solution

The user selects resources he/she wants to work with and defines their paths by means of the Preferences form in the software. It is thus possible to choose the Intex, Unitex and/or NooJ module, with a selected list of dictionaries. An important feature of WS4LR is its flexibility expressed both by the possibility of setting environment parameters and by the possibility of invoking command-line routines and using external Perl, Awk, and XSLT scripts. WS4LR functions and their usage are explained in a printed manual that accompanies the software, as well as in a concise on-line context sensitive help.

5. Conclusions

Although WS4LR has been used mainly for Serbian language resources, it is by no means language dependent. The only prerequisite is that the resources exist or are being developed according to the described formats and methodologies. Of course, not all of the resources need to exist. The user can work only on the resources he/she develops and modules that support them.

The development of WS4LR will continue as we intend to incorporate in it more sophisticated features. Namely:

- When new entries are added to a dictionary of compounds, it is presently the user that has to supply information on a compound constituent lemma, its inflectional code and grammatical categories. However, in most of the cases this information exists in DELAF type dictionaries of simple words, and we plan to make it available to the dictionary developer.
- Presently, a search key can only be a simple word lemma. We would like to enable a multi-word search as well, and to that end we plan to incorporate the multiword inflection module into WS4LR.
- Inflection for the target language in aligned texts is not yet supported. Namely, the translation equivalence option finds all synsets that contain the literals corresponding to the searching lemma in the wordnet of a source language and then the corresponding synsets in the target wordnet via the ILI. The search in the target language is then performed with synset literals only, without their inflected forms. We plan to include these forms in the search as well.

6. Bibliographical References

Courtois, B. & Silberztein, M. (eds.) (1990) Dictionnaires électroniques du français, Langue française 87, Paris: Larousse

Fellbaum, C. (ed.) (1998) WordNet: An Electronic Lexical Database, The MIT Press

Krstev, C., et al., (2004) Combining Heterogeneous Lexical Resources, in Proc. of the Fourth International Conference LREC, Lisbon, Portugal, May 2004, vol. 4, pp. 1103-1106

Horák, A., Smrž, P. (2004) New Features of Wordnet Editor VisDic, Romanian Journal of Information Science and Technology, Volume 7, Numbers 1-2, 2004, pp. 201-214

Savary, A. (2005) Towards a Formalism for the Computational Morphology of Multi-Word Units, in Proceedings of 2nd Language & Technology Conference, April 21-23, 2005, Poznan, Poland, ed. Zygmunt Vetulani, pp. 305-309, Wydawnictwo Poznanskie Sp. z o.o., Poznan

Stamou S., et al. (2002). BALKANET: A Multilingual Semantic Network for Balkan Languages, in Proc. of 1st International Wordnet Conference, Mysore, India

Veronis, J. (ed.) (2000) Parallel Text processing: Alignment and Use of Translation Corpora, Dordrecht: Kluwer Academic Publishers

Vossen, P. (ed.) (1998) EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Dordrecht: Kluwer Academic Publishers

Appendices

A. Morphological Dictionary Formats

DELAS

farmakolog,N12+Hum
dokolicyarenxe,N300+VN
Robinzon,N1002+Hum+NProp+First+Fict

DELAF

farmakolog,farmakolog.N+Hum:ms1v
farmakologa,farmakolog.N+Hum:ms2v:ms4v:mw2v:mw4v:mp2v
farmakologe,farmakolog.N+Hum:mp4v
farmakologom,farmakolog.N+Hum:ms6v
farmakologu,farmakolog.N+Hum:ms3v:ms7v
farmakolozi,farmakolog.N+Hum:mp1v:mp5v
farmakolozima,farmakolog.N+Hum:mp3v:mp6v:mp7v
farmakolozxe,farmakolog.N+Hum:ms5v

DELAC

vojno-tehnicyki(tehnicyki.A2:adms1g),
AC_2XA2//military technic
redovni(redovni.A2:adms1g) profesor(profesor.N2:ms1v),
NC_AXN+N+Comp//full-time professor

DELACF

redovna profesora,redovni profesor
.NC_AXN+N+Comp:mw2v
redovne profesore,redovni profesor
.NC_AXN+N+Comp:mp4v
redovni profesor,redovni profesor
.NC_AXN+N+Comp:ms1v
redovnih profesora,redovni profesor
.NC_AXN+N+Comp:mp2v

B. Wordnet Format

```
<SYNSET>
 <ID>ENG20-11902751-n</ID>
 <POS>n</POS>
 <SYNONYM>
  <LITERAL>pear<SENSE>2</SENSE>
  </LITERAL>
  <LITERAL>pear tree<SENSE>1</SENSE>
  </LITERAL>
  <LITERAL>Pyrus communis
       <SENSE>1</SENSE>
  </LITERAL>
 </SYNONYM>
 <ILR>ENG20-11902961-n
  <TYPE>hypernym</TYPE>
  </ILR>
 <ILR>ENG20-11902605-n
  <TYPE>holo_member</TYPE>
 </ILR>
 <DEF>Old World tree having sweet gritty-textured juicy
```

```
    fruit; widely cultivated in many varieties</DEF>
  <DOMAIN>botany</DOMAIN>
  <SUMO>FloweringPlant<TYPE>+</TYPE>
  </SUMO>
  <RILR>ENG20-07295527-n
    <TYPE>holo_part</TYPE>
  </RILR>
</SYNSET>

<SYNSET>
  <ID>ENG20-11902751-n</ID>
  <SYNONYM>
    <LITERAL>krusxka<SENSE>1</SENSE>
    </LITERAL>
    <LITERAL>Pyrus communis
      <SENSE>1</SENSE>
    </LITERAL>
  </SYNONYM>
  <DEF>Vocxka s glatkim listom i belil cvetovima, plodovi
su slatki i socyni sa karakteristicynim tvrdim
zrncima.</DEF>
  <POS>n</POS>
  <ILR>ENG20-11902961-n
    <TYPE>hypernym</TYPE>
  </ILR>
  <ILR>ENG20-11902605-n
    <TYPE>holo_member</TYPE>
  </ILR>
  <RILR>ENG20-07295527-n
    <TYPE>holo_part</TYPE>
  </RILR>
</SYNSET>
```

C. Format of Aligned Text

Serbian Original

```
<seg id="n94"> Sportska prognoza je igra u kojoj učesnik,
popunjavanjem listića koji izdaje priređivač igre na kojem
su označeni takmičarski parovi, pogađa rezultat fudbalske
ili druge sportske utakmice za svaki takmičarski par,
koristeći oznake predviđene pravilima igre. </seg>
```

English Translation

```
<seg id="n98"> Sports pool is a game in which a player
takes part by filling in a ticket, issued by the game
organizer, with previously printed opponents in matches,
e.g. soccer or other. </seg>

<seg id="n99"> The player guesses the results of the
matches on the ticket for each pair using symbols defined
by the rules of the game. </seg>
```

XML Output from XAlign

```
<link targets="n98 n99" type="linking" id="l3" />
<link targets="x94 l3" />
```

Aligned Text in TMX Format

```
<tu><prop type="Domain"> Files: 01Sreca-EN-id
01Sreca-SR-id</prop>
<tuv xml:lang="EN" creationid="n98 n99"
 creationdate="20040101T000000Z">
<seg>Sports pool is a game in which a player takes part
by filling in a ticket, issued by the game organizer, with
previously printed opponents in matches, e.g. soccer or
other. The player guesses the results of the matches on
the ticket for each pair using symbols defined by the rules
of the game.</seg> </tuv>
<tuv xml:lang="SR" creationid="n94"
 creationdate="20040101T000000Z">
<seg>Sportska prognoza je igra u kojoj učesnik,
popunjavanjem listića koji izdaje priređivač igre na kojem
su označeni takmičarski parovi, pogađa rezultat fudbalske
ili druge sportske utakmice za svaki takmičarski par,
koristeći oznake predviđene pravilima igre.</seg> </tuv>
</tu>
```

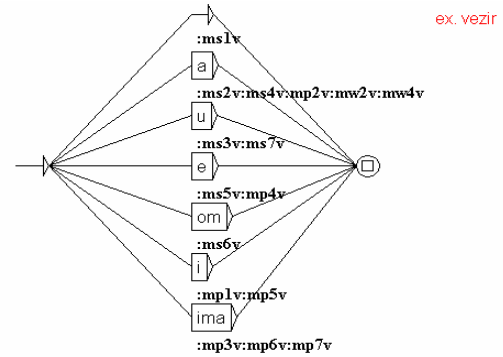

Figure 7: The inflectional transducer N2

D. Format of Transducers

```
10
"<E>" 100 200 7 2 3 4 5 6 7 8
"" 400 200 0
"<E>/:ms1v" 250 50 1 1
"a/:ms2v:ms4v:mp2v:mw2v:mw4v" 250 100 1 1
"u/:ms3v:ms7v" 250 150 1 1
"e/:ms5v:mp4v" 250 200 1 1
"om/:ms6v" 250 250 1 1
"i/:mp1v:mp5v" 250 300 1 1
"ima/:mp3v:mp6v:mp7v" 250 350 1 1
"ex. vezir" 467 53 0
```

E. Format of a Bilingual List

```
disjunction;disjunkcija
disjunctive normal form;disjunktivna normalna forma
disk;disk
disk cartridge;diskovna ulozxnica
disk drive;diskovni pogon
diskette;disketa
diskette drive;disketni pogon
disk file;diskovna teka
disk format;diskovni format
disk pack;diskovno paklo
disk unit;diskovna jedinica
```