Combining Heterogeneous Lexical Resources

Cvetana Krstev, professor, Faculty of Philology, Belgrade, cvetana@matf.bg.ac.yu
Duško Vitas, professor, Faculty of Mathematics, Belgrade, vitas@matf.bg.ac.yu
Ranka Stanković, assistant, Faculty of Mining and Geology, Đušina 7, Belgrade, ranka@rgf.bg.ac.yu
Ivan Obradović, professor, Faculty of Mining and Geology, Đušina 7, Belgrade, ivano@rgf.bg.ac.yu
Gordana Pavlović-Lažetić, professor, Faculty of Mathematics, Belgrade, gordana@matf.bg.ac.yu

Abstract

One of the main tasks of the Natural Language Processing Group at the Faculty of Mathematics, University of Belgrade is the development of various lexical resources. Among them the two most important ones are: the system of morphological dictionaries of Serbian (SMD) in Intex format and the Serbian wordnet (SWN) developed in the scope of the Balkanet project. Although these two resources represent dictionaries of a different type, developed using different models, each of them contains information that can either be incorporated into the other dictionary or that can be used in its development. In this paper we will outline some of the most interesting examples. We also present an integrated programming tool that enables the integration of these diverse lexical resources, as well as possible applications. We envisage the use of these resources in defining and linking lexical data in a way that will enable their more effective retrieval, integration, and reuse across various Web applications.

1 Introduction

One of the main tasks of the Natural Language Processing Group at the Faculty of Mathematics, University of Belgrade is the development of various lexical resources. Among them the two most important ones are:

- The system of morphological dictionaries of Serbian (SMD) in Intex format (Silberztein, 2000), that consists of a dictionary of simple lemmas, a dictionary of compounds (under construction), the corresponding dictionaries of word forms, and morphological finite-state automata that model certain classes of lemmas. The current size of SMD of simple lemmas is around 65.000, and they produce a dictionary of word forms with more than 930.000 entries. An example of an entry in the dictionary of simple lemmas (DELAS) is:

(1) devojcyin,A1+Pos+Ek

The information that has to be assigned to every entry is the part of speech (PoS) and the code of the inflectional class (for inflectional lemmas). Every inflectional class is implemented as a finite state transducer that is used to produce the dictionary of simple word forms (DELA F). Optional morphosyntactic, semantic and information on dialect can also be added. In the example (1), the lemma *devojcyin* belonging to a girl' is an adjective belonging to inflectional class A1. The adjective is possessive (+Pos), in ekavian pronunciation (+Ek). The information assigned to a lemma in DELAS can be used by Intex to formulate complex queries. For instance, the query <A+Pos-Ek> would retrieve all the possessive adjectives from a text that do not belong to the Ekavian pronunciation.

The entries in DELAS can be enriched by derivational links that group together entries belonging to the same derivational nest. This kind of information is given after an underscore sign. For instance,

(2) devojcyin,A1+Pos+Ek_N=4ka
devojka,N617+Hum+Ek_A=2cyin

The information in the first line states that the adjective *devojcyin* is linked to the noun entry and also indicates the way to identify this noun in the dictionary. Conversely, the information in the second line links the noun to the adjective. Moreover, the morphosyntactic information, preceded by a plus sign, can describe the type of derivational relation between two entries. In the example (2), the adjective *devojcyin* is the possessive adjective of the noun *devojka.* This information in the DELAS dictionary can be used by finite transducers to lemmatize the text using any lemma, arbitrarily chosen, from the derivational nest.

- The Serbian wordnet (SWN) is being developed in the scope of the Balkanet project (Stamou, 2002) following the model adopted for the EuroWordnet project (Vitas, 2003). The current size of SWN is 6290 synsets with 10583 literal string-sense pairs. Since the core of wordnets developed for Balkan languages was produced by translation of the basic synsets in the Princeton WordNet 2.0, the hypernym/hyponym relations in SWN mirror its hierarchical structure. Other relations are implemented more freely, depending on specific lexicalizations in Serbian. These relations include antonymy, meronymy, as well as some cross-part of speech relations (XPoS), such as CAUSES and BE_IN_STATE. For instance, synset [zatvoriti:1a] (synset [close:3, shut:1] in English WN (EWN), 'make shut') is in relation CAUSES with synset [zatvoriti se:1x] (in EWN [close:8, shut:2] 'become closed'). Also, synset [bolestan:1] (in EWN [unhealthy:1]) is in relation NEAR_ANTONYM with synset [zdrav:1] (in EWN [healthy:1]) and in relation BE_IN_STATE with synset [bolest:1x] (in EWN [illness:1, unwellness:1,...]).

All the WN developed in the scope of the Balkanet project use a common format for the exchange and linkage of data in XML. Although all BWNs use the same core XML schema, each wordnet can enhance this schema for some particular purposes (Figure 1).

Although these two resources represent dictionaries of a different type, developed using different models, each of them contains information that can either be incorporated into the other or be used in its development. In section 2 of this paper we will describe what information from one dictionary can be reused to the benefit of the other. In section 3 we will describe the software tool that performs these tasks, and some obtained results will be given in section 4.

2 The exchange of information

2.1 Usually, the only grammatical information accompanying the synset literal in a wordnet is the PoS, and it has to be the same for all literals in one synset. Enriching the synsets with information from SMD makes the usage of a wordnet in an information retrieval task more efficient. In a number of cases this additional information disambiguates the otherwise homonymous literals. This additional morphological, syntactic, and semantic information can be transferred from the SMD of simple forms and associated to each synset literal in SWN. For instance, in the following two synsets

(3) (**obaviti**:A1x, uraditi:4) (do:3, perform:4)
(okruzxiti:4, **obaviti**:B1v) (smother:1, surround:3)

the homographous literal **obaviti** appears. In both cases its is a verb but in two different inflectional paradigms (for the verb in the first synset the first person singular present form is *obavim* while for the verb in the second synset it is *obavijem*). Information about the inflectional properties can be found in the DELAS dictionary in a form of an inflectional class code. This information can thus be attached to each simple word literal string in WN, in a form of a XML element included into the literal element. For the example (3), the information attached to the literal **obaviti** from the first synset would be V157+Perf+Tr+Iref, while the same literal in the second synset would get the information V135+Perf+Tr+Iref. The additional morphosyntactic information states that in both cases the verbs are perfective, transitive and irreflexive. In some other cases, as for the synsets

(4) (**piti:1a**, popiti:4) (drink:1 , imbibe:3)
(**piti:1b**) (drink:5, tope:1)

one literal string, in this case the verb **piti**, represented in a traditional dictionary by one lemma, has two senses that bear different morphosyntactic features. The verb **piti** from the first synset ('take in liquids') would have the information V35+Imperf+Tr+Iref—transitive irreflexive progressive verb attached to it — while in the second synset the attached information would be V35+Imperf+It+Iref—itransitive progressive ireflexive verb.

For this kind of information the element <LNOTE> can be used that is contained in the element <LITERAL> in the core XML schema of the Balkanet project.

2.2 In order to overcome the restriction that only the same PoS literals can be part of one synset, XPoS links have been added to English and other wordnets, through relations such as CAUSES, BE_IN_STATE, DERIVED, PARTICLE, etc. The derivational information from the Serbian dictionary of simple forms can be used not only to add such and similar links to SWN but also to enrich it with synsets containing derived literals. For instance, the following five entries from the DELAS dictionary

(5) povezati,V122+Perf+Tr+Iref+Ref_V=3ivati_A=2n
povezivati,V18+Imperf+Tr+Iref+Ref_V=5ati_N=2nxe_A=2n
povezivanxe,N300+VN_V=3ti
povezan,A1+PP_V=1ti
povezivan,A1+PP_V=1ti

are derivationally connected, the derivational links being marked by underscores. Thus, the mark _V=3ivati attached to the first entry links the perfective verb *povezati* 'to join, to associate' to its corresponding progressive form *povezivati*, while the mark _A=2n links it to the adjective derived from its passive past participle *povezan*. The type of derivational link is marked by the plus sign—the progressive verb by the +Imperf mark and the passive past participle by the +PP mark. Similarly, the progressive verb *povezivati* in the second line of the example (5) is linked to its corresponding perfective verb *povezati*, the passive past participle *povezivan* and the verbal noun *povezivanxe*. This information can be used to link and/or add synsets containing such literals through the DERIVED relation. For example, the following two synsets in Serbian WN were linked by the DERIVED relation using this information:

(6) (zdruzxiti:1**, povezati:1**, ...) (join:2, bring together:1)
(**povezan:4**) (connected:2)

Many Serbian verbs can be used both as reflexive and irreflexive ones, which is marked by attaching both +Iref and +Ref marks to an entry in Serbian MD. SWN synsets containing such literals can potentially be related by CAUSES relation. The synsets presented in example (7) were linked by the CAUSES relation using this information.

(7) (odlomiti:1,..., **otkinuti:1a**) (chip:5,knap:2, cut off:3,...)
(**otkinuti se**:1b, odlomiti se:1) (chip:1. chip off: 1,...)

The relations introduced in this way sometimes mirror the existing relations in English WN, but they are more often specific to Serbian WN. One objection can be made to this kind of relations: they are derived from literals and not from synsets themselves. However, such relations exist in the Princeton WN as well, and they can be understood as 'at least on literal from the source synset is in the relation with al least one literal in the target synset'.

2.3 The information from SWN can be successfully used to enrich the SMD, namely the wordnet hierarchy can be used to add semantic information to simple word entries in SMD. Some basic semantic information has already been attached to simple word entries in Serbian MD, such as +Hum (human) and +Bot (botanic) for nouns, and +Col (colour) and +Mat (material) for adjectives. The use of wordnet enables a more systematic and more detailed attachment of such marks. Moreover, the attachment can be modeled according to the envisaged application. The hierarchy corresponding to the following branch in EWN

(8) abstraction:6
attribute:2
property:3
sound property:1
sound:1

can yield the addition of appropriate semantic marks to the entry *glas* 'voice' (hyponym of 'sound:1): glas,N16+Snd+SndProp+Prop+Attr+Abstr. Depending on the application the depth of the tree hierarchy and/or its level can be chosen.

Since only some basic semantic information has been incorporated in the Serbian MD, there are a number of identical entries—that is, apparently same lemmas with identical inflectional classes and morphosyntactic information attached to them—but actually representing different lemmas that can not be distinguished. That is the

case, for instance, with the double entry cyelo,N300 that represents both (brow:1, forehead:1) and (cello:1, violoncello:1). By adding the information obtained from the WN hypernym/hyponym relations these two entries can be distinguished: for instance, cyelo,N300+BodyPart and cyelo,N300+Artifact, or cyelo,N300+Thing+BodyPart+Feature and cyelo,N300+Artifact+Device+MusicInstr if more semantics is used.

Since a wordnet is not structured as a tree, but rather as an acyclic graph, it is possible that a synset has more than one hypernym, which allows multiple paths from a root node to such a synset. This is the case with the synset (show:3) 'public exhibition or entertainment':

(9) abstraction:6
relation:1
social relation:1 event:1
communication:2 social event:1
show:3

Lemmas from DELAS that correspond to literal strings of synsets that belong to the multiple hierarchical trees inherit semantic information from all those trees. For instance, the lemma corresponding to the literal string predstava:3a (equivalent of show:3) can inherit semantic information from both trees represented in (9) yielding the DELAS lemma predstava,N600+Abstraction+Event.

3 The supporting software

The C# programming language and Microsoft .Net were chosen as the basic development tools. Visual Studio .NET is a complete set of development tools for building ASP Web applications, XML Web services, desktop applications, and mobile applications. Visual Basic .NET, Visual C++ .NET, Visual C# .NET, and Visual J# .NET all use the same integrated development environment (IDE), which allows them to share tools and facilitates the creation of mixed-language solutions. In addition, these languages leverage the functionality of the .NET Framework, which provides access to key technologies that simplify the development of ASP Web applications and XML Web services.

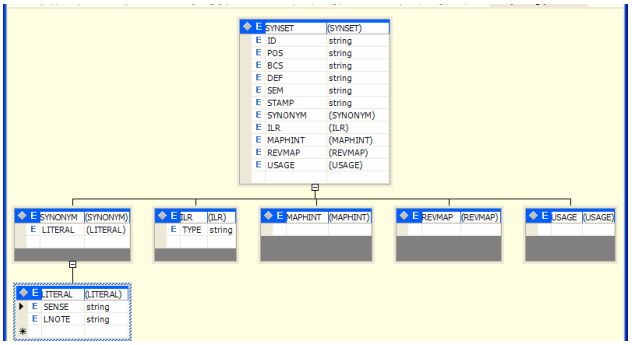

Figure 1 Serbian WN XSD schema in .NET XML designer. The element SEM is specific for SWN.

As XML is at the core of many features of Visual Studio .NET and the .NET Framework the existence of Serbian WN and other Balkan languages' WNs in XML exchangeable format facilitates their application. On the other hand, the XML Schema definition language (XSD) enables the definition of the structure and data types of XML documents. Figure 1 shows the graphical representation of XSD schema of Serbian WN.

The XML Path Language (XPath) provides a language for addressing parts of an XML document. XPath treats an XML document as a tree of interrelated branches and nodes. A node in a XML document can be an element, attribute, processing instruction, comment, textual content, namespace, and document itself. The XPath tree model is based not on the nodes themselves, but rather on their mutual relationship. For example, the way elements relate to one another, the way attributes relate to elements, and so on. For instance, the following XPath expression

(10) //SYNSET[POS='n' and not(ILR/TYPE='hypernym')]

retrieves from a XML document representing wordnet using the XSD from the Fig 1 all the nouns without hypernyms, i.e. first in a hierarchy.

In this environment an Integrated Language Resource Management Tool (ILReMaT) has been developed. This tool supports the development of a wordnet in accordance with the BWN, and enables its integration with other lexical resources. This tool has been designed as a complement to the VisDic software, a tool accepted by all participants in the Balkanet project for wordnet development. For instance, it detects discrepancies between the hypernym/hyponym trees in two wordnets, synsets from a chosen subset of base concepts that have not yet been included in the wordnet under development, synsets which lack some information (for instance, glosses) etc. Besides that, this tool enables the integration of a with bilingual word list in the development process, which can help in the translation of literal strings and in the checking of the existing cross-language relations.

Various editing tasks can be performed on a synset, which can be chosen in several ways: simply by typing one of its literal strings, or by selecting it by means of the bilingual list, by following the hypernym/hyponym or some other relation from the current working synset, or even by typing one's own XPath expression. An edit form is provided for the working synset in which the content of all elements can be filled and updated. More than one edit form can be opened at each moment, thus enabling easy updates of related synsets. In this form one field corresponds to the element SEM that is specific to the Serbian WN. This element contains semantic information designating the synset concept. This information can then be added using the plus mark to all the lemmas in DELAS that correspond to the concept's hyponyms.

The main feature of ILReMaT is its capability to work with a wordnet and DELAS dictionaries in parallel and to enable transfer of information from one type of a resource to the other. In order to perform these tasks special tab-pages of the edit form are designed. For instance, the tab-page "Intex graph" enables the production of a graph for a working synset that can be used in Intex the environment to retrieve all the inflective forms of all the synset's literal strings.

The other two tab-pages are aimed at performing the tasks described in section 2. The tab-page "Update with Intex dictionary" enables the inclusion of morphosyntactic information from Serbian MD into a working synset. An element LNOTE, which is in the content of the LITERAL element in the XSD schema common to all Balkan languages. is used in the Serbian WN for morphosyntactic information specific for this literal. This information is automatically retrieved from the DELAS dictionaries. If more than one instance is retrieved from these dictionaries, the user can choose the

appropriate one. Moreover, he can modify (delete or add) the automatically retrieved information.

Once the lemmas corresponding to a synset are retrieved from the DELAS dictionaries, the ones that have derivational information attached to them (marked with an underscore) can be used for retrieving synsets from the WN that contain literals derived from those lemmas. The user can then choose to establish a derivational relation from the original synset to some of the retrieved ones, where appropriate. The program automatically infers the type of derivational relation from the derivational marks in the working and retrieved synsets.

The second tab-page "Intex SEM" is used for retrieving semantic information that is going to become the content of SEM element, from all the working synset's hypernyms up to the root element. Then all the lemmas corresponding to working synset's literals are retrieved from the DELAS dictionary, and a string of plus sign marks is formed, which the user can choose to add to the retrieved DELAS lemmas. If due to the addition of semantic information one DELAS lemma has to be separated in two or more lemmas, the copies of the original lemma can be made and appropriate semantic information added to each of them.

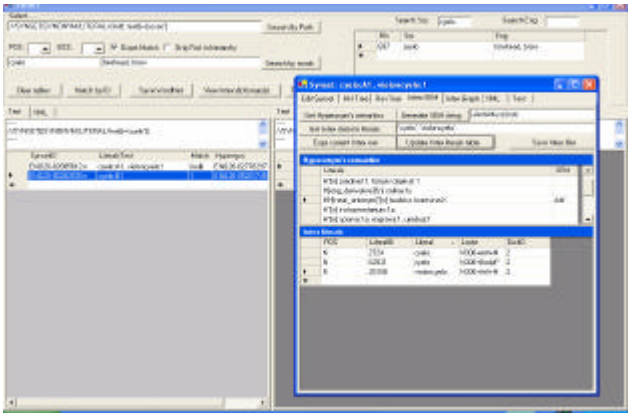

Figure 2. The tab-page "Intex SEM" shows the separation of the lemma cyelo.

The ILReMaT tool has many more additional features — we have presented her only the ones dealing with the integration of lexical resources have been presented.

4 Some Results

The developed tool has been applied for the purpose of enhancing of both the Serbian WN and the DELAS dictionaries. Morphosyntactic information has been added to literals of most of the synsets in Serbian WN. Our aim is to provide this information for all simple literals in Serbian WN by the end of the Balkanet project. That would enable the correct production of all their inflected forms, a task that has been planned for all the languages that are participating in the project.

At this moment only some information related to the structural derivation has been included in the DELAS dictionaries, such as the derivation of verbal nouns from verbs or possessive adjectives from nouns. Thus, appropriate relations were introduced specific to the Serbian WN, such as DERIVE_VN and DERIVE_POSA, and they have been applied using ILReMaT on the working version of WN. For instance, the synset (asociranxe:2, povezivanxe:6) (association:5, connection:5, connexion:3) is related with the synset (povezati:6) (associate:1, tie in:2, relate:1,…) by the relation DERIVE_VN. It should be noted that these two synsets are also related by the relation ENG_DERIVATIVE. This, however, is not the case in general.

The semantic marks in the DELAS entries enable the formulation of complex queries in the Intex environment. The use of the Intex regular expression <N+MusicInst> (noun marked as a musical instrument in any form) enables, for instance, retrieving from a text of all the phrases of the type "to play on a musical instrument".

Da bi na gitari moglo da se svira solo dodata je josx jedna zxic muzike iz Vaca, koji na klaviru i fruli sviraju baroknu muziku, Mnogi gitaristi koji sviraju klasicynu gitaru kada uzmu u ruke ovu

The inclusion of these semantic signs in DELAS dictionary the whole funcionality of WN can not be achived. They, however, contribute to the functionality of Serbian MD.

5 Conclusion

The process described in this paper has proven beneficial for both kinds of the resources. However, we should point some problems in its application. First of all, the sizes of Serbian MD and Serbian WN are not comparable. The development of Serbian MD has started good many years before WN, so it more thoroughly covers the language. As a consequence, the Serbian MD can benefit less form the WN then vice versa. For that reason, the production of the fully semantically marked Serbian DELAS has been postponed until the two resources will become comparable in size.

Besides that, the development of the Serbian morphological dictionary of compounds is in its initial phase, which is a serious drawback for the enhancement of the WN with morphosyntactic information, where a number of literals are compounds, e.g. (usmeni ispit:1) 'oral exam', with their own inflectional rules. This problem will be gradually resolved as the Serbian MD of compounds grows.

Finally, the developed tool does not perform any of the tasks automatically, although that solution was also under consideration. Since the Serbian traditional lexical resources can not be directly used for the production of electronic resources, and almost none exist in electronic form, the Serbian resources presented in this paper have been manually produced, checked and double checked. Our standpoint is that only when reliable lexical resources in electronic form are fully developed it will be possible to produce new resources automatically.

Bibliography

- Silberztein, M. (2000). *INTEX Manual*, Paris: Asstril.
- Stamou S., et al.: (2002). BALKANET: A Multilingual Semantic Network for Balkan Languages. Proceedings of 1st International Wordnet Conference, Mysore, India.
- Vitas, D. et al. (2003). Resources and Basic Tools for the Processing of Serbian Written Texts. Proc. of the Workshop on Balkan Language Resources, 1st Balkan Conference in Informatics.
- Vossen, P. (ed.) (1998). EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Dordrecht: Kluwer Academic Publishers.