Local Grammars and Compound Verb Lemmatization in Serbo-Croatian

Introduction

The increasing production of electronic (digital) texts (either on the Web or in other electronically available forms, such as digital libraries or archives) demands appropriate computer tools that can help human users in text manipulation and, additionally, in performing automatic processing of language resources. In the first place, a natural language processing (NLP) system needs to implement models for recognition and isolation of various lexical constituents that occur in a digital text. The main purpose is to isolate and mark syntactical units for further analysis. In this paper we will analyze a method for modelling compound verb forms in Serbo-Croatian, and, consequently, an approach to automatic recognition of string sequences in a digital text which represent such forms.

NLP processing of a text in a highly inflective language needs to include thorough lexical preprocessing and lemmatization, that must take into account various inflected forms of words. The goal of lemmatization is to determine a lemma for each textual word as well as the appropriate grammatical information that corresponds to it. For example, the process of lemmatization for nouns usually assigns to a textual word its nominative form and corresponding grammatical information (e.g. gender, number, case and other properties). The main objective of the paper is to analyze methods for lemmatization of compound verb forms that occur in a digital text. However, as parts of compound verb forms can be locally distributed within a sentence, the lemmatization has to “bring together” parts of the same verb form and to determine the lemma (i.e. the infinitive form of the verb) and corresponding grammatical information (e.g. gender, person, tense). Unlike the lemmatization of nouns and adjectives, which has been exhaustively studied (cf. [3] and [12]), the references on (compound) verb lemmatization are not numerous, even for the languages that are not highly inflective. As a starting point for our research we have used a framework exposed in [2], dealing with compound forms of verbs in English.

The structure of the paper is as follows. First, we will briefly discuss problems of compound verb lemmatization and the resources and tools that we used. Afterwards, we will present a case study of lemmatization of the preterite tense in Serbo-Croatian and some other examples of verb phrase lemmatization.

Lemmatization and Compound Verb Forms

Verbs in Serbo-Croatian (SC), as in many other languages, are conjugated by means of suffixes and by using auxiliary verbs. Conjugated forms of a verb are built basically in two different ways: verb forms are either simple (e.g. the Present Tense in SC: On peva ‘He sings’) or compound — verb form consists of several lexical units, usually an auxiliary and a main verb (e.g. the Preterite Tense in SC: On je pevao ‘He was singing’). Both simple and compound verb forms are parts of verb paradigm, so it is justified to include both of them in a process of lemmatization.

The model for building a compound verb form is usually described in detail in traditional grammar books. For example, a verb form in the Preterite in SC is built using the present tense of the auxiliary verb jesam ‘to be’ and the active participle (or so-called active verbal adjective) of a main verb, both conjugated in the appropriate gender and number (e.g. On je pevao ‘He was singing’, One su pevale ‘They were singing’). Information presented in grammar books, however, is usually inadequate for automated processing of text. For example, grammar books, as a rule, mention standard orderings of the constituents, while exploration of corpora reveals various additional examples of compound verb form usage. In order to perform automated lemmatization of compound verb forms, we need to take into
account possible structures that match “well-formed” compound verb forms. Here, we will briefly present the problems (exposed in detail in [2]) of textual dispersion of lexical units that form a compound verb, which are to be resolved during the process of compound verb lemmatization:

1. word order

The phenomenon of pseudo-free word order is characteristic for Slavic languages, and for SC as well. An auxiliary verb can precede a main verb, but a main verb can also be ahead of an auxiliary. Consider examples:

\[\ldots\text{samo je znao nesxto o cyemu}\ldots\]
\[\ldots\text{i znao je mnoge vesxtine}\ldots\]

The meanings are similar, but not the same: the underlying sequence of elements is semantically relevant.

2. inserts

Inserts of various types are allowed between the constituents of a verb form. Usually, the inserts are different kinds of adverbial and/or noun phrases:

\[\ldots\text{koje je u potpunosti potisnulo skromne hauzmajstore}\ldots\]
\[\ldots\text{zbog cyega je gospodja dolazila}\ldots\]

Inserts can be as simple as one word (for example, an adverb or a noun), but can be embedded sentences as well. If, however, all these inserts are correctly parsed, practically all compound verb forms can be recognized [2].

3. omissions

Some parts of a compound verb form can be omitted (usually an auxiliary verb):

\[\text{Taj me miris dotakao kada sam,\ldots}\]

In such situations, it is sometimes difficult to determine some of the corresponding grammatical information: in some contexts the grammatical codes can be traced, but this requires for a thorough analysis of the wider context.

As indicated before, the goal of the paper is to define a model and a set of procedures for automatic recognition of compound verb forms, regardless of the distribution of their constituents in a digital text. To accomplish the task, we have to use exhaustive linguistic resources (already available for computer processing) and appropriate tools for the analysis. The next section summarizes the resources that were used.

---

1 Examples presented in this paper are taken from the novel “Lagum” by Svetlana Velmar-Jankovic. To neutralize the problem of Serbo-Croatian dual alphabet (Cyrillic and Latin), we used the encoding format in which specific SC characters are encoded by two characters (e.g. cy, cx, sx, zx, lx denote č, ě, š, ž, lj respectively). Consequently, a text encoded by means of this format is independent of the alphabet, with which it was originally recorded. The tools presented in the paper use this encoding format as well.
Resources and Tools

We can systematize language resources that we have used as follows.

a) *Traditional grammar books* are used to define the basic structures of language phenomena. Usually, traditional grammars are neither complete nor sufficient for formal description of a phenomenon, and, especially, for implementation of computer procedures. They are created for human use, and not for automatic processing. Along with grammar books, paper dictionaries represent language resources with similar purpose and scope. These resources cannot be used for computer processing without thorough examination.

b) *Electronic dictionaries* are morphological dictionaries of a special form suitable for computer processing. Unlike paper dictionaries, electronic dictionaries (e-dictionaries) are not designed for human usage. They contain exhaustive description of morphosyntactic characteristics of a specified language. The system of electronic dictionaries has been described in detail in [3], [12] and [14]. Along with nouns, adjectives and adverbs, e-dictionaries contain entries for verbs. Verbs are classified according to a type of conjugation that is performed in order to obtain their paradigms [5]. For each verb in the DELAS dictionary, its complete paradigm is given in the DELAF dictionary. Figure 1 illustrates the paradigm of the verb *dolaziti* ‘to come’:

```plaintext
DELAS     DELAF

dolaze,dolaziti,V:P3p
dolazecxi,dolaziti,V:AdvPr
dolazi,dolaziti,V:P3s:Y2s:A2
dolazicx,dolaziti,V:P2s
dolazicxe,dolaziti,V:F3s:F3p
dolazicxemo,dolaziti,V:F1p

dolaziti,V34.07.4*

dolazicxesx,dolaziti,V:F2s
dolaziti,dolaziti,V:W
dolazivsx,dolaziti,V:AdvPs
dolaxzh,dolaziti,V:11s
dolaxxwu,dolaziti,V:13p
```

Figure 1: The excerpt from DELAS/DELAF dictionaries for SC

E-dictionaries are used for lexical recognition and initial lemmatization of words that occur in a digital text. This process is realized by dictionary look-ups [10], which result in an *initially tagged* text: each textual word is associated with its lemma and corresponding morphosyntactic categories (tags) stored in e-dictionary. In general, e-dictionaries cannot resolve lexical ambiguities resulting from the fact that there is no one-to-one correspondence between word forms and their morphosyntactic categories. There are methods to suppress ambiguities by using the cache-dictionaries [11] in order to force some look-ups and to ignore others.

c) *Local grammars* are finite-state transducers (cf. [3], [4] and [11]) that are used to recognize "well formed" word sequences in a text and to choose the proper tags for them. A local grammar describes a set of related strings that have a similar structure and that can be processed in a uniform way. It does not describe a sentence as a whole; however, it defines and applies lexical constraints to a local environment containing a sequence of several words. A local grammar can be represented either by a graph or by a regular expression, and it is applied to a digital text after the initial tagging. For example, the local grammar
pored.PREP <N:g>

expresses the following constraint: a noun that follows the preposition pored ‘by’ has to be in genitive case and, therefore, other possible tags of the noun, if any, can be removed. If the noun form is not in genitive case, than the sequence is considered as ungrammatical. It is essential for a local grammar neither to give incorrect tags nor to remove any possible lexical interpretation that can be applied in the local environment. Similarly, the local grammar:

koliko <PRO:a> grlo nosi

describes a frozen adverbial expression (‘as loud as one can shout’) in which the pronoun (<PRO>) has to be in accusative case.

Local grammars can be defined to follow the rules stated in traditional grammar books, but some can be added after the corpus analysis [8]. They can be used to model complex lexical and syntactical phenomena. For example, local grammars can be used:

- for lemmatization and lexical disambiguation of textual words (e.g. discarding morpho-syntactic information that is not applicable in the local context [7]);
- for recognition of syntactical units (e.g. noun phrases (cf. [6]), adverbial phrases, dates, etc.);
- as transducers (e.g. translation from one dialect of the language into another).

d) A corpus can be used both as a source for research and as a test area for defined models and procedures [8]. In our research, we have used newspaper texts and various literary texts.

For processing the resources mentioned, we have used the system INTEX – an integrated corpus processing system developed at the LADL (Laboratoire d'Automatique Documentaire et Linguistique), Université Paris VII (cf. [10], [11]). The system integrates the support for:

- initial tagging of a digital text using the system of e-dictionaries;
- defining and applying local grammars to initially tagged text;
- locating patterns described by a regular expression or by a graph;
- corpus analyzing.

Local Grammars for Compound Tense Recognition

In this section we will illustrate the process of constructing local grammars for compound verb lemmatization using the preterite tense as an example. The proposed method, however, can be implemented in the same manner for other compound tenses as well.

The very first attempt to lemmatize the compound verb forms was a simple extraction of sequences of strings that form the compound verb in the preterite, ignoring the problems mentioned in the second section. The underlying structure of the compound verb form corresponded to the following regular expression:

\[ <jesam.V:P> (<MOT>)* <V:PP> + <V:PP> (<MOT>)* <jesam.V> \]

The application of this local grammar to some sample digital texts has resulted in severe noise: since the inserts were not parsed, the set of recognized strings was too broad, i.e. the grammar was “over-fitting”. Moreover, this grammar did not model agreements between

---

2 The expression \(<V:P>\) denotes a verb form in present, while \(<V:PP>\) denotes the active participle of a verb. The expression \(<MOT>*\) denotes any sequence of words (including an empty sequence).
constituents. Therefore, the problem of recovering complete information corresponding to a verb form remained.

In order to obtain more accurate and reliable structures, the corpus was consulted using the system INTEX. We have extracted context (in form of concordances) of active participles of verbs. By analyzing the concordances, the structures from corpus were manually compiled into appropriate local grammars. The process of local grammar tuning by means of corpus consulting was iterative and—as a result—the number of structures and corresponding local grammars has significantly increased, while the noise has decreased. As opposed to the methods applied for extractions of prepositional phrases [8] and compound names [9], where we have mostly used lexical constraints, in this approach we have focussed on structures of sequences that are to be recognized. To obtain local grammars which are both reasonably complex and readable (especially for the sake of presentation and maintenance), we have divided them into separate subgraphs (e.g. a subgraph for each person; for instance, ten subgraphs were constructed for the preterite tense). Here we present the “general” local grammar for the preterite tense (figure 2), and its subgraphs for 3rd person singular, feminine (figure 3) and for 1st person plural, masculine and feminine (figure 4).

Figure 2: A general local grammar for the preterite tense

The local grammar in figure 3 represents the possible ordering of a main verb and an auxiliary, as well as the distribution of inserts. For example, three branches of the graph in figure 3 describe the following positions of the constituents in a verb form:

\[ NP \ Aux \ Inserts \ Verb \] \ or \ \[ Aux \ NP \ Inserts \ Verb \] \ or \ \[ Verb \ Aux \]

\[ 3 \] Shaded boxes denote other graphs and subgraphs that are automatically called by the INTEX parsing program.
Inserts that we have considered are either nouns (in the appropriate gender and person), or some adverbial phrases represented by the subgraph $Adv$. Since there is a shortage of a systematic research into the syntax of adverbial phrases in SC, we have defined the subgraph $Adv$ containing adverbs (stored in the e-dictionary of SC) and restricted list of adverbial phrases (e.g. $s\ vremena\ na\ vreme$ ‘from time to time’, $u\ potpunosti$ ‘completely’).

Figure 3: A local grammar for the preterite tense, 3rd person singular, feminine
Figure 4: A local grammar for the preterite tense, 1st person plural

For the purpose of lemmatization, some constituents are enclosed with enumerated brackets. These constituents can later be referred to as ‘$’ variables (cf. the output of the node for $V:PP$ in figure 3). The lemmatization is performed by means of output of the local grammars, i.e. the tags assigned during the initial tagging are replaced with corresponding output. A form of an auxiliary verb jesam ‘to be’ is replaced by a tag <AUX> (or by ne-<AUX> for the negation) to keep track of its position, and a main verb tag is augmented with codes for the appropriate tense (e.g. $PT$ denotes preterite tense), person and gender (e.g. $sf$ denotes singular, feminine).

Local grammars that have been defined can be applied to a digital text either to obtain lemmatized concordances (e.g. to locate all patterns that are recognized as preterite verb forms), or to transform the input text by tagging the sequences that are recognized as compound verb forms. The figures 5, 6 and 7 present a few examples of results of local grammar application.

Figure 5 shows the excerpt from the sorted concordances of strings recognized by the local grammar for the preterite tense (figure 2). After applying local grammars to a digital text, the lemmatized concordances can be generated by the system INTEX. Figure 6 further presents a few samples from the same text: each sample contains a sequence which occurred in a text (the first lines in the figure), the initially tagged sequence (the second lines), and the lemmatized sequence (the third lines). In the first example, a noun $gospodja$ ‘lady’ appears as an insert; in the second and third examples, adverbs (a compound adverb $u$ $potpunosti$ ‘completely’ and a simple adverb $visxe$ ‘more’) are inserts. All these inserts were correctly parsed by the local grammar.

| nastojnici | <AUX> {{bili,biti.V:PPpm}.PT3pm} tu, u tim stanovima |
| rat taj hol | <AUX> {{bio,biti.V:PPsm}.PT3sm} prekriven crvenom sagom a po |
| z bog cyeega | <AUX> gospodja {{dolazila,dolaziti.V:PPsf:PPpn}.PT3sn}, |
| koji | <AUX> {{hteo,hteti.V:PPsm}.PT3sm}, da izgleda strog |
| Tih godina | <AUX> {{izbegavala,izbegavati.V:PPsf:PPpn}.PT1sf}, da izlazim, |
| Sa Saborne | <AUX> {{izbijalo,izbijati.V:PPsn}.PT3sn}, podne, zenit |
| tek sxto | <AUX> {{izbio,izbiti.V:PPsm}.PT3sm} rat, mozxda vecx |
| a Jermenin | <AUX> {{izigledao,izgledati.V:PPsm}.PT3sm} obicyno. |
| isti kamioni | <AUX> {{odonosili,odonosit.V:VPpm}.PT3pm}, i lexseve, naslagane |
| otkud zna | {{pitala,pitati.V:PPsf:PPpn}.PT1sf} <AUX> se, da je |
| crnu kosu. | {{pitao,pitati.V:PPsn}.PT3sn} <AUX> sxapatom skliskim, |
| josx bila tu | {{podrhtavala,podrhtavati.V:PPsf:PPpn}.PT1sf} <AUX> u bakalnici |
| Uskrsa | {{raspale,raspasti.V:PPpf}.PT3pf} <AUX> se ne samo kucxe |
| toliko starao, | {{raspali,raspasti.V:PPpm}.PT3pm} <AUX> se u tren oka i isxciyezli u |
| polica i izloga, | {{stajale,stajati.V:PPpf}.PT3pf} <AUX> velike, |
| iznad nxe, | {{stajali,stajati.V:PPpm}.PT3pm} <AUX> neobicyno cyisti dyakov |
| po kvasac, | {{vratila,vratiti.V:PPsf}.PT1sf} <AUX> se sa strepxnom koja, sada |

Figure 5: An excerpt of lemmatized concordances of verbs in the preterite tense

| cxega | je gospodja dolazila |
| cxega | <AUX> {gospodja,.N:fsn+:fpg+} {{dolazila,dolaziti.V:PPsf:PPpn}.PT3sf} |

| koje | je u potpunosti potisnulo skromne |
| koje | {{je,jesam.V:VP3s}.{u potpunosti,.ADV}{{potisnulo,potisnuti.V:PPsn}} skromne |
| koje | <AUX> {u potpunosti,.ADV} {{potisnulo,potisnuti.V:PPsn}.PT3sn} skromne |
Figure 6: Samples of the process of compound verb form lemmatization

Figure 7 shows a sample of tagging verb forms in a text. A textual form of a verb is associated with the appropriate lemma (the infinitive form of the verb) and complete grammatical information. Note that after this phase of lemmatization, the textual words denoting compound tenses were disambiguated.

Figure 7: Compound verb forms tagged in a digital text

In addition to the usage of local grammars for compound tense lemmatization, they can be used for recognition of other verb phrases, for instance, recognition of negations associated with simple verb forms or for recognition of modal verb forms (ex. *Voleo bih da pevam* ‘I would like to sing’).

Conclusion and Further Research

This paper presents the initial stage of research in using local grammars for modelling verb phrase recognition in a highly inflective language such SC, by exemplifying recognition of compound verb forms. The research has made it clear that the level of formal description of the language (both its quantity and quality) determines the possibilities for automatic text manipulation. Formal lexical and syntactical data related to a linguistic phenomenon are essential for its processing and handling.

The model which describes compound verb forms and their lemmatization is illustrated using the preterite tense as an example, but similar models can be designed for other compound tenses (and for other linguistic phenomena, as well). The model takes into account some of the problems presented in the second section (ordering of the constituents, lexical distribution, different kinds of inserts). However, the corresponding local grammars are by no means complete, although they do not make false recognitions. By examination of large corpora and by completing e-dictionaries, local grammars can be improved and tuned (cf. [8] and [9]).

The representation of local grammars by graphs appears to be “well-adapted to modelling syntax of natural languages” [2]. Further, local grammars provide a computationally effective means of recognition of lexical and syntactical structures in a digital text, and, hence, large quantities of texts can be processed efficiently in real time. The applications of the local
grammars presented here include lexical disambiguation, generation of lemmatized concordances, tagging of compound verb forms as syntactical units, improved indexing, etc.

For further research we plan to enlarge and complete the local grammars that have been presented, and to test them on a wider corpus. New contexts and new constraints, that are likely to appear, will be built into grammars. This will, consequently, upgrade the coverage of recognition of compound verb forms in a digital text. The complete parsing of inserts and embedded structures is another problem for further research. In addition, processing of omissions has not yet been considered, as some cross-references through a text have to be done in order to resolve the problem.

References


