SENTENCE CLASSIFICATION AND CLAUSE DETECTION FOR CROATIAN

Kristina Vučković*, Željko Agić*, Marko Tadić**

*Department of Information Sciences, **Department of Linguistics
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, 10000 Zagreb, Croatia
{kvuckovi, zeljko.agic, marko.tadic}@ffzg.hr

ABSTRACT

We present a method for classifying Croatian sentences by structure and detecting independent and dependent clauses within these sentences and provide its evaluation. A prototype system applying the method was implemented by using the NooJ linguistic development environment, both for purposes of this experiment and for further utilization in a prototype rule-based chunking and shallow parsing system for Croatian. With regards to pre-processing, we implemented and evaluated three different approaches to designing the system: (1) no pre-processing of input sentences, (2) automatic morphosyntactic tagging of sentences by using the CroTag stochastic tagger and (3) manual morphosyntactic annotation of input sentences. All three approaches were evaluated for sentence classification and clause detection accuracy in terms of precision and recall. The highest scoring system was the one using sentences with manually assigned morphosyntactic tags as input: an overall F1-measure of 0.861 (P: 0.928, R: 0.813). In the paper, a more detailed discussion of system design and experiment setup is provided, followed by a discussion of the obtained results and future research directions.

1. Introduction

Many natural language processing tools and complex natural language processing systems assembled by pipelining these tools demand certain methods of pre-processing the text input in order to operate at required levels of accuracy and efficiency or more generally, from a software engineering point of view, in order to meet the various functional and non-functional user requirements. One of the basic pre-processing tasks in language technologies is the segmentation of input text into paragraphs, sentences, tokens, etc. Here, we inspect the problem of sentence segmentation from a less common viewpoint. The problem of sentence segmentation – separating the input text into sentences – is well known to be resolved by using simple regular expressions with an accuracy of above 99 percent correctly detected sentence boundaries. However, building on top of e.g. (Boras 1998), we choose to inspect Croatian sentences from a more elaborate – both linguistically and computationally motivated – perspective. Namely, we do not seek solely to detect the sentence boundaries, i.e. to discover beginnings and endings of sentences, but also to (1) detect boundaries of clauses in complex sentences and (2) detect their type according to the grammatical classification of Croatian sentences. In this way, an implementation of such an analysis of input text written in Croatian may at the same time serve various linguistically motivated inquiries and also be used as a pre-processing module for more complex pipelined natural language processing systems, the latter being further elaborated in the following sections of the paper.

In Croatian, we classify sentences by their purpose or by their structure (Barić et al. 2005). By purpose, we differentiate between declarative, interrogative and exclamatory sentences. By structure, we classify the sentences into two major groups: (1) simple sentences and (2) complex sentences. Simple sentences contain only one predicate, paired with a subject only or with both a subject and object(s). Complex sentences are subcategorized into (1) independent complex, or compound sentences and (2) dependent complex sentences. There are six types of compound sentences where the independent clause is connected to the main clause by using a conjunction and being coordinated with the main clause. Based on the conjunction used, we can differentiate the following types of dependencies between the clauses: (1) constituent clauses, conjunctions {*i, pa, te, ni, niti*}; (2) disjunctive clauses, conjunction {*ili,ili…ili*}; (3) opposite clauses, conjunctions {*a, ali, nego, no, već*}; (4) exclusive clauses, conjunctions {*samo, samo što, jedino, jedino što, tek*}; (5) conclusive clauses, conjunctions {*dakle, zato, stoga*} and (6) explanatory clauses, conjunctions {*jer, ta, jerbo*}. Dependent complex sentences are made by connecting two or more clauses in such a manner in which the main clause is independent while all the other clauses depend on the main clause and cannot stand alone in a sentence. Namely, we distinguish predicate, subject, object, attribute, apposition and adverbial type of clauses.

This research is based on and developed keeping in mind the specific requirements of systems that have already been developed for parsing simple Croatian sentences consisting of a subject, verb, direct and indirect object, adverbial of time, place and manner (Vučković et al. 2008, Vučković 2009). With this research we hope to extend the existing model to recognize the before-mentioned sentence structures, which includes both classifying sentences by structure and detecting clauses within them in order to parse Croatian sentences more efficiently.

In the following section, a system – implemented as a NooJ module – for sentence and clause detection and classification in Croatian texts is presented. Sections 3 and 4 present the experiment's plan and discuss the obtained results. We conclude by sketching possible future research directions, specifically in terms of pipelining the system presented here to the chunker and partial parser for Croatian (Vučković et al. 2008, Vučković et al. 2009, Vučković 2009, Vučković et al. 2010).

2. Detection and classification system

The sentence and clause detection and classification system for Croatian is implemented in the latest version of the NooJ linguistic development environment (Silberztein 2003, Silberztein 2008, Silberztein 2009, Silberztein 2010).

The main grammar for Croatian clause detection consists of two types of sub-graphs: main clauses (*Mainclause* and *Mainclause2*) and independent clauses (*IndependentClauses1*, *IndependentClauses2*) that may appear once before or any number of times after the main clause(s). The main grammar is displayed in Figure 1, as shown in NooJ.

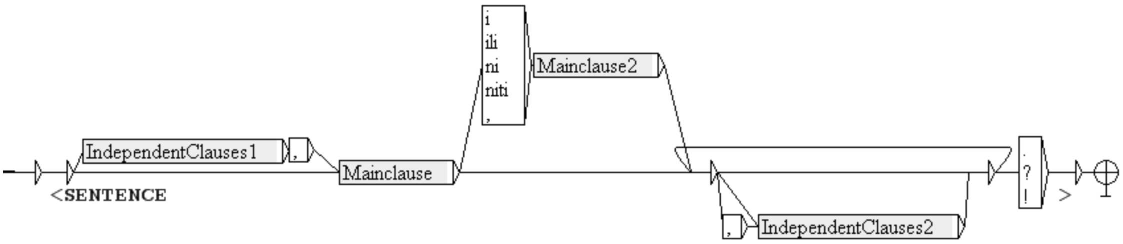

Figure 1. Top-level grammar of the system

As indicated by Figure 2, the main idea behind the grammar for clause detection is the presence of only one verb <V> or verb phrase <VP> and possibly any other phrase (noun phrase <NP> left part of Figure 3, prepositional phrase <PP>, adverb <R>, conjunction <C>, numeral <M>, pronoun <PRO>, preposition <S>) including the brackets and quotation marks as shown on the right part of Figure 3. This is true for all types of clauses (dependent and independent ones). The main difference between these two types of clauses is that the independent clauses do not depend on the verb from the main clause and may be recognized only according to the conjunction that they start with (see Figure 4). This is, however, not true of the dependant clauses with the exception of Adverbial dependant clause type. In order for the dependant clauses to be recognized, more than the conjunction recognition is necessary.

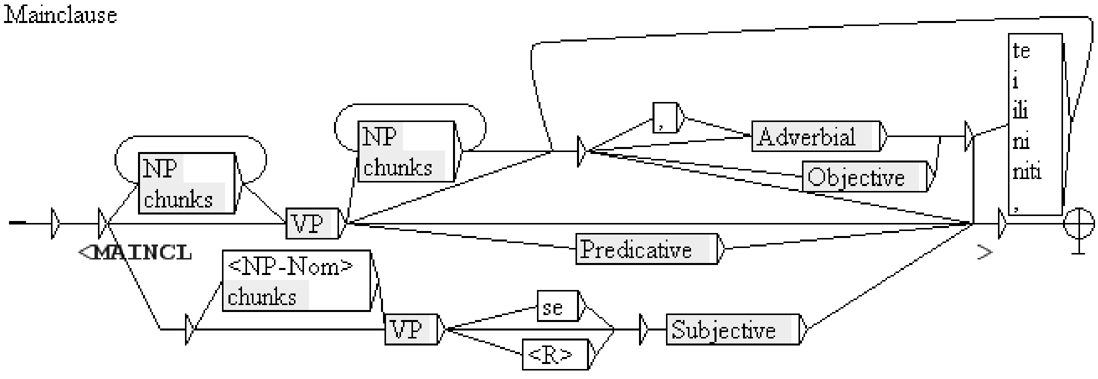

Figure 2. *Mainclause* grammar

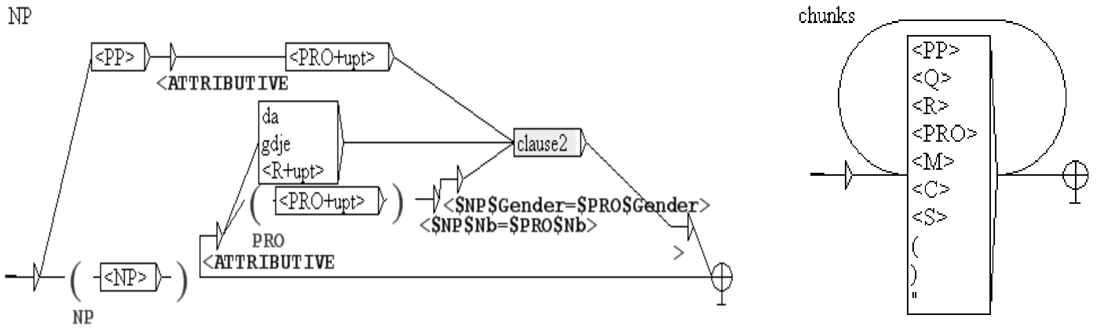

Figure 3. Grammars for detecting noun phrases and other chunks

Object dependant clause grammar has to check if the verb from the main clause takes as a complement an accusative noun phrase <NP+Acc>. Only then it continues to check if the type of the conjunction ({*da, gdje, kako, kuda, neka, kamo, koliko, kad*} or an interrogative pronoun) characteristic for this type of clauses is present before entering the clause itself.

Predicate dependant clause follows immediately after the verb from the main sentence and its grammar first checks if the verb from the main clause is any form of an auxiliary verb *to be* (hr. *biti*) or its negation *not to be* (hr. pres. 1p sg. *nisam*) since the entire predicate clause behaves as a nominal predicate to that verb. If this condition is met, the grammar checks if the conjunctions are of the predicate type ({*takav da, takva da, takvo da, tolik da, tolika da, toliko da*} or an interrogative pronoun) before entering the sub-graph for the clause recognition.

Subject dependant clause has a prerequisite of different form than the two previously explained clauses. It requires that any <NP> that may be present in the main sentence must not be in Nominative case <NP-Nom> since this entire clause behaves as a subject of the main clause. If this condition is met and if the conjunction of the clause is any from the following set ({*da, gdje, kako, kamo*} or an interrogative pronoun), the grammar will proceed to the clause itself.

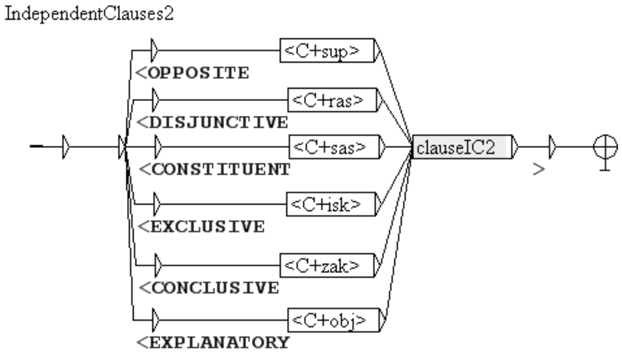

Figure 4. *IndependentClauses2* grammar

Adverbial dependant clause does not require any additional checks and it depends only on the conjunction that precedes it. The choice of a conjunction defines also the subtype of an adverbial clause in the following manner:

- Adverbial of time - {*kad, kada, dok, dokle, dočim, čim, jedva, tek, netom, pošto, kako, što, otkada, otkako, poslije nego, prije nego, kad god, dok god, dokle god, sve dok, samo dok, jedva što, tek što, istom što, netom što, nakon što, poslije nego što,pošto, prije nego što*}

- Adverbial of manner - {*kako, kao što, kao da, koliko, što, kano da, kanda*}
- Adverbial of place - {*gdje, kamo, kud, kuda, otkud, otkuda, odakle, dokle*}
- Adverbial of cause - {*jer, što, kad, kada, kako, budući da, jerbo, gdje, zašto, zato što, stoga što, zbog toga što, uslijed toga što, zahvaljujući tomu što*}
- Adverbial of condition - {(*i*) *ako, ukoliko, samo ako, samo da, samo kad*}

Attribute dependant clause may appear after any noun in the sentence. This is the reason why it is described together with the noun phrase <NP> (Vučković et al. 2008, Vučković 2009, Vučković et al. 2010). However, since the prepositional phrase <PP> ends with an <NP> chunk, it was necessary to add it into this sub-graph as well. Thus, if there is a <PP> followed by the interrogative pronoun, this pronoun may open the place for an entire clause. The similar goes for the noun phrase. If there is an <NP> followed by *da*, *gdje*, any interrogative type of adverb or any interrogative pronoun, this may open the place for an attribute clause. The difference between <NP> and <PP> being followed by a pronoun, is that in the first case number and gender agreement between the <NP> and a pronoun have to be checked (see Figure 3). However, this is, at the moment, not possible to check in the case of a <PP> chunk.

3. Experiment setup

At the time of conducting the experiment, no gold standard corpus containing the information on sentence types and clause boundaries was available. The Croatia Weekly 100 kw (CW100) corpus (cf. Tadić 2002, Vučković et al. 2008), being manually morphosyntactically annotated, lemmatized and XCES-encoded up to and including the sentence level, contained information on sentence boundaries. Thus, the CW100 corpus was used to manually assemble a small gold standard for the purposes of this specific experiment. More precisely, 200 sentences were chosen from the sentence pool of the CW100 corpus – 100 were assigned for the development stage, used for designing the system itself and other 100 were used for evaluation purposes. The sentences were chosen randomly, but the randomization process itself was biased towards selecting the sentences of average length, the average for CW100 being approximately 25 tokens, thus avoiding too short or too long sentences.

The detection and classification system was used in three different settings. In the first one, the 100 sentences of the testing set were provided as an input for the system without pre-processing. The second run used the CroTag stochastic morphosyntactic tagger (cf. Agić et al. 2008) to pre-process the sentences, with tagger accuracy of approximately 85 percent correctly annotated tokens, while the third run contained 100 percent accurate morphosyntactic annotation coupled with the sentences, as the annotation taken from the CW100 metadata for the sentences of the testing set. The outputs of these three systems were then evaluated and the results are discussed in the following section.

4. Results and discussion

Three sets of observations on the performance of the systems were made on their respective outputs. In this section, they are represented by tables 1, 2 and 3.

The first table gives an insight into the overall performance of the three systems in terms of their precision, recall and F1-measure as observed on the testing set. Recall was measured as a ratio of detected and existing clauses in the test sentences and it shows the system using manually tagged sentences (or an ideal tagger as a pre-processing module) as the top-performer with recall of 0.813, substantially higher than the other two systems. Interestingly enough, the table also shows that the noise introduced by the CroTag tagger and its decrease in tagging accuracy when compared with the ideal tagger actually decreases the detection performance even below the baseline set by the first system, i.e. the one using no pre-processing at all. However, when measuring system precision, utilizing the CroTag tagger does, in fact, somewhat improve the results and precision rises steadily from the first to the third system, even though this is not manifested in the respective F1-measures for the three systems, being that their differences for recall are more substantial than for precision.

	No tagging	**CroTag**	**Manual tagging**
Existing	289	289	289
Detected	211	190	235
Recall	0.730	0.657	**0.813**
Classified	187	169	218
Misclassified	23	19	17
Precision	0.890	0.899	**0.928**
F1-measure	0.802	0.759	**0.867**

Table 1. Precision and recall of the systems

Table 2 provides a type distribution for correctly detected classified sentence clauses across the systems. The table serves both as an indicator of the actual distribution of clauses in the testing set and as an addition to the general scores given in the previous table. Excluding the main clauses (MAIN), being that they expectedly dominate the distribution, object (OBJ), opposite (OPP), adverbial (ADV_CAUSE and others) and attribute (ATT) clauses are the most frequently encountered ones.

Sentence type	**No tagging**	**CroTag**	**Manual tagging**
ADV_CAUSE	14	12	16
ADV_COND	2	3	3
ADV_CONSEQ	0	0	1
ADV_MANNER	0	1	2
ADV_PLACE	0	0	1
ADV_TIME	2	1	2
ATT	10	8	15
CONST	10	9	9
DISJUNCT	1	1	1
EXPL	0	1	1
MAIN	85	79	95
OBJ	21	16	20
OPP	39	35	48
SUB	3	3	4

Table 2. Distribution of correctly classified sentences

Table 3 is basically a sentence classification confusion matrix, given summary for the three systems. The top row of the table is an indicator of the actual classification while the leftmost column lists all the misclassifications that occurred within the testing sample. As an illustration, the object clause (OBJ) was misclassified as main clause (MAIN) three times. Closely correlating with Table 2, attribute (ATT) and object (OBJ) clauses are most commonly misclassified, even though a larger testing set would surely provide a more informative confusion matrix.

	ADV_CAUSE	**ADV_CONSEQ**	**ADV_MANNER**	**ADV_PURPOSE**	**ATT**	**CONSEQ**	**CONST**	**MAIN**	**OBJ**	**OPP**	**PURP**	**SUB**
ADV	0	0	0	0	0	0	**2**	0	**2**	0	0	**2**
ADV_CAUSE	0	0	**1**	0	0	0	0	0	**1**	0	0	0
ADV_TIME	0	0	0	0	0	0	0	0	0	**1**	0	0
ATT	0	0	0	0	0	0	0	0	**3**	0	0	0
CONCL	0	**3**	0	0	0	0	0	**1**	0	0	0	0
CONST	0	**1**	0	0	0	**1**	0	0	0	0	0	0
EXPL	**2**	0	0	0	0	0	0	0	0	0	0	0
MAIN	**2**	0	0	0	**3**	0	**2**	0	**3**	**7**	0	0
OBJ	0	0	0	0	**2**	0	0	0	0	0	0	**1**
PRED	0	0	0	**3**	0	0	0	0	0	0	0	**1**
SUB	0	0	0	**3**	**4**	0	0	0	**5**	0	**3**	0

Table 3. Confusion matrix for sentence classification

Figure 5 shows disambiguation in the main clause and the independent opposite clause. They both have a dependant clause that can be interpreted fourfold as an object dependant clause or as an adverbial dependant clause of cause, manner or time. This ambiguity will be marked for all clauses that start with the conjunction *kako* if the verb from the main clause is a transitive verb that takes noun phrase in accusative as its complement. If the verb is intransitive, the clause will still be marked ambiguously as an adverbial clause of cause, manner or time only not as an object clause. There are more similar disambiguation problems due to the fact that different types of clauses share the same conjunctions.

Coordination of verbs is a problem that still needs to be revisited, especially in models where there are compound verb forms present in the sentence but the auxiliary verb "to be" is present for only one occurrence of the verb (usually the first one), while it is only implicitly transferred to the other verbs in a sentence. For example, in sentences such as: *Marko je pjevao u siječnju i glumio u travnju* (en. *Marko was singing in January and acting in April*). Nominal predicate also presents a problem at the present that will need to be looked more deeply into. For exmaple, such is a sentence: *Zakon bi nudio povlastice i ako je investicija usmjerena na ona područja Hrvatske koja su sada po gospodarskoj snazi ispod državnog prosjeka.* (en. *The law would offer benefits even if the investment is directed towards those areas of Croatia which are now concerning the economic power below the state percentage.*) with three clauses and two of which have a dislocated nominal predicates ('je …usmjerena', 'su… ispod').

```
<sentence>
  <maincl>Mesić je kazao
    <adverbial type="cause"><adverbial type="manner"><adverbial type="time">
          <objective>kako ne želi biti fikus</objective>
    </adverbial></adverbial></adverbial>
  </maincl>
  <opposite>dok je Račan ponovio
    <objective>
      <adverbial type="cause"><adverbial type="manner">
        <adverbial type="time">kako su se građani Hrvatske opredijelili za parlamentarnu demokraciju
      </adverbial></adverbial></adverbial>
    </objective>
  </opposite>.
</sentence>
```

```
<sentence>
  <maincl>Ivo Škrabalo je kasnije kazao
    <adverbial type="cause"><adverbial type="manner"><adverbial type="time">
          <objective>kako je Vijeće HRT-a jednoglasno dalo potporu programu direktora Galića</objective>
    </adverbial></adverbial></adverbial>
  </maincl>
  …
</sentence>
```

```
<sentence>
  …
  <opposite>a istaknuo je
    <subjective>
      <objective>
        <adverbial type="cause"><adverbial type="manner"><adverbial type="time">
          kako se radi i na donošenju Zakona o porezu na dohodak i Zakona o porezu na dobit
        </adverbial></adverbial></adverbial>
      </objective>
    </subjective>
  </opposite>.
</sentence>
```

Figure 5. Example sentences and annotation

Another problem is recognizing attribute dependant clause that starts with a prepositional phrase inside which there is an interrogative pronoun like '*za koje*' in the following example: *… jer je ponudio program promjena* ***za koje*** *su potrebni novi ljudi, …* (en. *… since he had offered a program of changes* ***for which*** *new people are needed, …*).

5. Conclusions and future work

We presented and evaluated a rule-based module developed in NooJ for sentence and clause classification and detection in Croatian texts. Assembling the module with the CroTag morphosyntactic tagger, we created three systems and evaluated them for precision and recall. The top performing system, i.e. the one using ideal morphosyntactic annotations for the input sentences, reached the F1-measure of 0.861 (precision: 0.928 recall: 0.813).

Further improvements of the system itself might include solving the problems with coordination of verbs, nominal predicates and attribute clases that start with a prepositional phrase (as explained in the previous section) but also detection of dependant clauses in other positions like when inserted between the main and its auxiliary verb or insertion of dependant clauses deeper into the sentence structure (beyond level 3 insertion that the system recognizes now) and combination of direct and indirect speech clauses. We are also planning an experiment with linking the system presented here with the rule-based chunker and shallow parser for Croatian (Vučković et al. 2008, Vučković et al. 2009, Vučković 2009, Vučković et al. 2010) in order to improve its overall accuracy on Croatian texts.

Acknowledgements

The research within the project ACCURAT leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement no 248347. This work was also supported by the Ministry of Science, Education and Sports, Republic of Croatia, under the grants No. 130-1300646-1776 and 130-1300646-0645.

References

Agić Željko, Tadić Marko, Dovedan Zdravko. (2008). Improving Part-of-Speech Tagging Accuracy for Croatian by Morphological Analysis. Informatica 32:4, pp. 445-451.

Barić Eugenija, Lončarić Mijo, Malić Dragica, Pavešić Slavko, Peti Mirko, Zečević Vesna, Znika Marija. (2005). Hrvatska gramatika, Školska knjiga, Zagreb.

Boras Damir. (1998). Teorija i pravila segmentacije teksta na hrvatskom jeziku. PhD Thesis, Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, 1998.

Silberztein Max. (2003). NooJ Manual. http://www.nooj4nlp.net/NooJ Manual.pdf, 2003.

Silberztein Max. (2008). Complex Annotations with NooJ. Proceedings of the 2007 International NooJ Conference. Cambridge Scholars Publishing, Newcastle, pp. 214-227.

Silberztein Max. (2009). Syntactic parsing with NooJ. Finite State Language Engineering: NooJ 2009 International Conference and Workshop, Centre de Publication Universitaire, pp. 177-189.

Silberztein Max. (2010). Disambiguation Tools for NooJ. Proceedings of the 2008 International NooJ Conference. Cambridge Scholars Publishing, Newcastle (in press).

Tadić Marko. (2002). Building the Croatian National Corpus. Proceedings of the 3rd International Conference on Language Resources and Evaluation, ELRA.

Vučković Kristina, Tadić Marko, Dovedan Zdravko. (2008). Rule Based Chunker for Croatian. Proceedings of the Sixth International Conference on Language Resources and Evaluation, Paris-Marrakech, 2008.

Vučković Kristina. Model parsera za hrvatski jezik. (2009). PhD Thesis, Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, 2009.

Vučković Kristina, Bekavac Božo, Dovedan Zdravko. (2009). SynCro – Parsing simple Croatian sentences. Finite State Language Engineering: NooJ 2009 International Conference and Workshop, Centre de Publication Universitaire, pp. 207-217.

Vučković Kristina, Agić Željko, Tadić Marko. (2010). Improving Chunking Accuracy on Croatian Texts by Morphosyntactic Tagging. Proceedings of the Seventh International Conference on Language Resources and Evaluation, European Language Resources Association, Valletta, pp. 1944-1949.

Vučković Kristina, Tadić Marko, Bekavac Božo. (2010). Croatian Language Resources for NooJ. Proceedings of the 32nd International Conference on Information Technology Interfaces, SRCE University Computer Centre, University of Zagreb, Zagreb, pp. 121-126.