Coll. Antropol. **27** Suppl. 1 (2003) 195–198
UDC 808.62-3:519.6
Original scientific paper

Database of Sentences of Croatian Language

Božidar Tepeš and Vladimir Mateljan

Department of Information Sciences, Faculty of Philosophy, University of Zagreb, Zagreb, Croatia

A B S T R A C T

The aim of this paper is to investigate morphological and syntactical levels of sentences of Croatian Language. Morphological grammatical level is represented by 16 categories and 47 features with theirs value. Syntactical level is represented with constituent structure trees of sentences. Database has 1001 grammatical sentences of Croatian literature. Database of sentences is the result of theoretical research in computational linguistics. This study extends the analysis of the linguistic data in anthropology in Croatian. Access to database is through the Internet on the address: http://infoz.ffzg.hr/tepes.

Key words: *linguistic database, probabilistic context-free grammar, Croatian language.*

Introduction

A long-term anthropological research of population structure of the East Adriatic rural populations on a number of Adriatic islands estimate basic geographical, historical, economic, demographic and linguistic factors. Linguistic factors directly or indirectly might have influenced the formation of the island population structure[1–8]. The method of hidden Markov model was aimed at identification of internal and external impulse of change or continuity of rural populations within a wider socio-cultural context[9–11]. This studies now been extended to investigation on morphological and syntactical structure of Croatian language. Sentences in the database of sentences of Croatian language are from literature[12]. The project with large language database is project of Linguistic Data Consortium[13]. The theoretical background is theory of formal and natural language[14].

Probabilistic context-free grammar

Context-free grammar[15] is formal system *CFG(V,T,S,P)*, where V is finite set of variables, T is finite set of terminals, $S \in V$

Received for publication April 13, 2003

is start symbol and P is set of production $A \to w$ for $A \in V$ and string $w \in (V \cup T)^*$. Production define production relation |– between strings:

$$(w_1 \mid\!- w_2) \Leftrightarrow ((w_1 = \alpha A \beta) \wedge (w_2 = \alpha w \beta) \wedge$$

$$((A \to w) \in P))$$

Reflexive and transitive closure of production relation is |= derivation. Context-free language *CFL* is set of strings of terminals that are derived from start symbol:

$$CFL = \{x \mid (S \models x) \wedge (x \in T^*)\}$$

Context-free grammar *CFG* is deterministic context-free grammar.

Probabilistic context-free grammar[16–17] is a formal system $PCFG(CFG(V,T,S,P),p)$ where $CFG(V,T,S,P)$ is deterministic context-free grammar and p is a set of discrete probability distributions of productions that have same variable A on the left hand side of productions $A \to w$. Discrete distributions have to satisfy the norm condition:

$$\sum_{\substack{(A \to w) \in P \\ w \in (V \cup T)^*}} p(A \to w) = 1$$

Head-driven phrase structure grammar

Natural language has syntax that is described with models of generative grammar, lexical functional grammar and head -driven phrase structure grammar.

The minimalist program of generative grammar[18], describes the natural language by using lexicon, computational system and constituent structure tree. Lexicon is described as a set of lexicon items together with its features. Main lexical features are categorial features, ϕ-features, case features and strong features. Categorial features are connected with syntax and semantic of lexical items. Computational system takes lexical items from lexicon and forms new syntactic objects. Result of sequence of operations in computational system is structural descriptions of phrases and sentences. Structural description is constituent structure tree.

In case of lexical functional grammar[19] constituent structure tree is just a part of sentence's description, called external structure. Every nod of constituent structure tree has a function. Function f maps attributes of node ATT in values of attributes *VAL*:

$$f: ATT \to VAL$$

Head-driven phrase structure grammar[20] describes lexical items with the feature structure. Each feature has its value of the feature. Features, together with the values, make the feature structure. Main lexical features are syntactic and semantic features. The most important syntactic features are head category, specifier and complement. Principles and rules of the head-driven phrase structure grammar connect lexical items in sentences.

Database of sentences of Croatian language

In the database of sentences of Croatian language each sentence has constituent structure tree. Each lexical item in sentence has its own features with features value. Constituent structure tree of sentence of writer V. Nazor, »Near the top of the hill stands rock.« (»Pri vrhu brda stoji stijena.«) is on Figure 1.

In this sentence the features of words and its values are:

»Pri« is part of speech: preposition,
»vrhu« is part of speech: noun; case: locative; number: singular; gender: masculine,
»brda« is part of speech: noun; case: genitive; number: singular; gender: neuter,
»stoji« is part of speech: verb; affirmation/negation: affirmative; voice: active; finiteness/infiniteness: finite: composi-

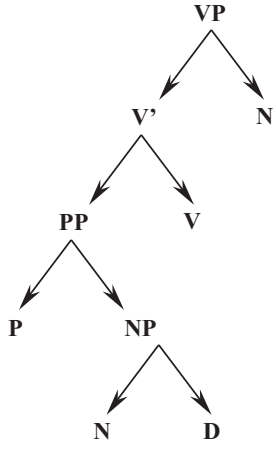

Fig. 1. Constituent structure tree.

tion: simple; tense: present; person: third; number: singular; aspect: imperfective, »stijena« is part of speech: noun; case: nominative; number: singular; gender: feminine.

In the database of sentences of Croatian language the constituent structure tree in Figure 1. is written in one row:

$$VP[V'[PP[P,NP[N,D]],V'],N]$$

Access and search of the database

Database of sentences of Croatian language is designed for determination of distribution as probabilities of probabilistic context-free grammar. Database enables testing of the theoretical models of the grammar for the Croatian language and their evaluation. main goals of making of the database of sentences of Croatian language are:

– Creation of database of annotated sentences of Croatian language;

– Creation of database, based on theoretical foundations of the computational linguistics;

– Creation of model of the automatic annotation of the sentences of Croatian language;

– To extends the analysis of the linguistic data in anthropology in Croatian.

Access to the database of sentences of Croatian language is through the internet, which technology requires standardization of annotation and knowledge of HTML and XML technologies. Creation of models of automatic annotation of sentences of Croatian language requires determination of parsers of Croatian language, which use probabilistic context -free grammars and hidden Markov models[21].

Database for testing the sentences of Croatian language is created on SQL server 2000. Database contains table of sentences, table of words with foreign key to the sentence to which the word belongs and table of features of words with foreign key to the word to which the features belonging. As more then one word belongs to each sentence and more then one features belong to each word, tables are in relation one to many. Each sentence is described by a sequence of words, constituent structure tree and syntactical and morphological features with values. Referential integrity is imposed between tables which, together with the protection of database on the level of SQL server and on the level of operative system (Windows 2000 server), enables maximum security and truthfulness of data in the base. Program for filling of the base has been made as well as ASP pages for its search. Base is implemented on the WEB server of Department of information sciences, faculty of philosophy, University of Zagreb and can be searched on internet by using address: http://infoz.ffzg.hr/tepes.

Acknowledgements

The research is funded by the Ministry of Science and Technology of the Republic of Croatia under grant 0130015 for the project »Tagging and Word Recognition for the Croatian Language«.

REFERENCES

1. RUDAN, P., D. F. ROBERTS, A. SUJOLDŽIĆ, B. MACAROL, ŽUŠKIN, A. KAŠTELAN, Coll. Antropol., 6 (1982) 39. — 2. RUDAN, P., D. ŠIMIĆ, N. SMOLEJ NARANČIĆ, L. A. BENNETT, B. JANIĆIJEVIĆ, V. JOVANOVIĆ, M. F. LETHBRIDGE, J. MILIČIĆ, D. F. ROBERTS, A. SUJOLDŽIĆ, L. SZIROVICZA, Am. J. Phys. Anthropol., 74 (1987) 417. — 3. RUDAN, P., A. CHAVENTRE, Coll. Antropol., 13 (1989) 177. — 4. SUJOLDŽIĆ, A., Coll. Antropol., 13 (1989) 189. — 5. SUJOLDŽIĆ, A., Coll. Antropol., 17 (1993) 17. — 6. SUJOLDŽIĆ, A., P. RUDAN, V. JOVANOVIĆ, B. JANIĆIJEVIĆ, A. CHAVENTRE, Coll. Antropol., 11 (1987) 181. — 7. SUJOLDŽIĆ, A., P. ŠIMUNOVIĆ, B. FINKA, L. A. BENNETT, J. L. ANGEL, P. RUDAN, Antropol. Linguistics, 28 (1987) 405. — 8. ŠKREBLIN, L., L. ŠIMIČIĆ, A. SUJOLDŽIĆ, Coll. Antropol., 26 (2002) 333. — 9. TEPEŠ B., T. ŽUBRINIĆ, L. SZIROVICZA, I. HUNJET, In: Proceedings. (Int. Conf. Information Technology Interface, ITI'96, 1996) — 10. SZIROVITCZA, L., A. SUJOLDŽIĆ, B. TEPEŠ, Coll. Antropol., 21 (1997) 609 — 11. TEPEŠ, B., L. SZIROVICZA, A. SUJOLDŽIĆ, M. PRIMORAC, Journal of Computing and Information Technology, 5 (1997) 265 — 12. KATIČIĆ, R.: Sintaksa hrvatskoga književnog jezika. (HAZU, Zagreb, 1991). — 13. BIAS, A., M. FERGUSUN, K. KATZ, R. MACINTURE: Bracketing guidelines for treebank II style Penn treebank project. (Linguistic Data Consortium, University of Pennsylvania 1995). — 14. TEPEŠ, B.: Računarska ligvistika. (University of Zagreb, Zagreb 2001). — 15. HOPCROFT, J. E., J. D. ULLMAN: Introduction to automata theory, languages and computation. (Addison-Weseley Publ. Comp., Reading, 1979). — 16. CHI, Z., S. GERMAN S. Computational Linguistics, 24 (1998) 299. — 17. CHI, Z., Computational Linguistics, 25 (1999) 131. — 18. CHOMSKY, N.: The minimalist program. (The MIT Press, Cambridge, Mass., 1996). — 19. BRESNAN, J.: Lexical-functional grammar. (Backwell Publishers, 2000). — 20. SAG, A., T. WASOW: Syntactic theory. (CLS Publications, Stanford 1999). — 21. STOLKE, A., Computational Linguistic, 21 (1995) 165.

B. Tepeš

Department of Information Sciences, Faculty of Philosophy, University of Zagreb, I. Lučića 3, 10000 Zagreb, Croatia

BAZA PODATAKA REČENICA HRVATSKOGA JEZIKA

SAŽETAK

Cilj je ovoga rada istraživanje morfoloških i sintaktičkih razina rečenica hrvatskoga jezika. Morfološka gramatička razina predstavljena je sa 16 kategorija i 47 obilježja. Sintaktička gramatička razina predstavljena je sa sastavnicama strukture rečenice. Baza podataka ima 1001 gramatičku rečenicu hrvatske književnosti. Baza podataka rezultat je teorijskog istraživanja u računarskoj lingvistici. Ova studija nastavlja analizu lingvističkih podataka u antropologiji u Hrvatskoj. Pristup bazi je preko Interneta na adresi: http://infoz.hr/tepes.