Automatic Morphological Generation and Analysis for the Croatian Language: Lexical Inflectional Database as Personal Name Recognition Module

Damir Boras, Davor Lauc, Nives Mikelić
Faculty of Philosophy, University of Zagreb
I . Lučića 3, Zagreb Croatia
e-mail: {dboras, dlauc, nmikelic}@ffzg.hr

Abstract: *This paper describes methodology for automatic morphological generation and analysis using inflectional lexical database. The database contains all existing personal names today in Croatia. It is possible to generate all word-forms for given names in accordance with the Croatian language rules. Rules for combination of personal names with family names are given. As personal name recognition module the system is included in robust morphological processor for the Croatian Language.*

Keywords: lexical database, inflected language, Croatian personal names, personal name recognition, morphological processing

1. Introduction:

There are different approaches dealing with problem of automatic analysis and generation of word-forms. Many morphological processors are based on two level Koskenniemi's model [3]. As a general computational model it has many advantages. But, for adaptation of two-level model to an inflected language such as Croatian, a special lexicon input module is needed. There is also a very simple approach based on robust morphologicl processing [4]. It is possible to derive all word-forms without a dictionary using all the theoretically possible combinations of morphonological alternations, and filtering a final results through a list of exisitng words. Furthermore, it is possible to create a lexical database for an inflected language such as Croatian that would derive all word-forms from some basic forms [1,2]. In this paper, the methodology for automatic morphological generation and analysis using lexical database of Croatian personal names is described.

In information retrieval context being able to retrieve on names, whether personal, institutional, geographic or other names, is an very important capability. Some applications use name searching to extend the traditional information retrieval paradigm [6]. Alhough there are aproaches dealing with NLP techniques, it is possible to apply inflectional database of personal names in determining whether a name of interest in a query matches a name in a textual database, in order to support free text retrieval.

Name searching can be defined as the process of using a name as part of a query in order to retrieve information associated with the name in a database. Name searching, in the general case, includes both name-recognition and name-matching. If names are not already identified as such in database's text records, e.g. when they appear as part of a free text field and have not been previously tagged as being names, name recognition is recquired. Similarly in parsing a query, if name has not been identified as name by query's syntax, then it will be necessary to recognize it. Once a name is recognized in a query and database record, then name matching algoritms are needed to determine whether the names are the same, or that thay in a fact designate the same individual.

So, recognizing strings being names is the first step of a name searching process. As a personal name recognition module it is possible to apply a database of names. Since Croatian is highly inflected language there are different word-forms of a name concerning different cases. So, to be efficiently used as a personal name recognition modul a database should comprise all existing names today in Croatia in all different word-forms. In this paper the inflectional lexical databes containing personal names is described. The database enable to derive all word-forms from a basic forms. Basic forms create the list of nominative case word-forms and to them a paradigms of word endings are associtated. Linguistically, the presented methodology allows to operate easy with a languages comprising reach morphology such as the Croatian language.

2. Characteristic of the Croatian language

Croatian is highly inflected language and there are fourteen different word-forms for nouns including singular and plural (ten in the written language, when accents are not marked). Also, there are more than hundred word-forms for adjectives (14 or 15 in the written language) taking into account definite and indefinite forms of adjectives, as well as more than twenty word-forms for verbs in simple tenses (more then hundred for compound tenses).

Croatian morphology of personal names deals with three genders, seven cases, tree types of declension and there are rules for combining personal names with family names.

Thre geneders are:

- masculine (m) - males
- feminine (f) - females
- neuter (n)

Seven cases are:

- ***Nominative*** – designating the subject of a finite verb
- ***Genitive*** – typically expressing possession, source or a partitive concept
- ***Dative*** – designating indirect object of a finite verb
- ***Accusative*** – designating the direct object of a finite verb
- ***Vocative*** – used for direct addressing
- ***Locative*** – indicating the place at which or in which
- ***Instrumental*** – expressing means or agency

Three types of declension are:

- ***First or a-declension*** – for almost all masculine and all neuter nouns (examples: Marko, Hrvoje)
- ***Second or e-declension*** – for feminine nouns and some masculine nouns (examples: Nikola, Kate)
- ***Third or i-declension*** – for feminine nouns ending with consonant (there is no personal names using this type of declension)

Rules for combining personal names with family names are following:

- Every female and male first name has its type of declension.

For example:

	male	**male**	**female**	**female**
N	Ivo	Nikola	Marica	Ivana
G	Ive	Nikole	Marice	Ivane
D	Ivi	Nikoli	Marici	Ivana
A	Ivu	Nikolu	Maricu	Ivanu
V	Ivo	Nikola	Marice	Ivana
L	Ivi	Nikoli	Marici	Ivani
I	Ivom	Nikolom	Maricom	Ivanom

- Personal names can also have plural forms.

- Vocative of the personal name can have form different from the nominative, but one can allways use the nominative form instead of vocative form (if the vocative

form is different from the nominative form) esspecially in the spoken or non-formal language.

Also, when using vocative form with family names addressing male persons, different usage has "social" meaning.

For example:

If you address "mister Marinović" you can use two different forms:

"gospodine Marinović" – when addressing formally

"gospodine Marinović**u**" – only if you are very good friend of mister Marinović, or it could be thought that you are adressing him ironically.

- Family names have declension only when related to the male persons.
- Family names related to female persons are indeclinable.

For example:

	male *name and fam. name* **female** *name and fam. name*
N	Ivan Botica Ivana Botica
G	Ivana Botić Ivane Botica
D	Ivanu Botić Ivani Botica
A	Ivana Botića Ivanu Botica
V	Ivane Botica Ivana (Ivano) Botica
L	Ivanu Botić Ivani Botica
I	Ivanom Boticom Ivanom Botica

The only exception to this rule is the female name combined with female second name acting as a family name of the patronymic type originated from Slavic languages (Russian, Macedonian, etc.).

For example.:

N Irina Rodnjina Jadranka Stojanovska

D Irini Rodnjinoj Jadranki Stojanovskoj

3. Croatian Personal Name Inflectional Database

The database contains all exisitng names today in Croatia collected from different sources such as all publicly available data bases (telephone directories, court registers for buiness companies etc.)

The database structure is determined names and paradigms (models for name inflection). Names are contained in the three tables: male first names, female first names, family names. Names table structure is given by following fields: name, paradigm tag – unique paradigm tag, alternative paradigm tag – tag for the alternative paradigm if it is exists and frequency (number of occurences in source data).

Names table

Ime	Cest	Prdg	Prdg1
damil	2	d	
damir	13650	d	

Paradigm table structure is given by following fields: mnemonic name – common name for every type of declension according to the name gender and ending, paradigm tag – unique paradigme tag, gender (**m,** males - masculine, "muška", **p,** - family names "prezimena", **z,** females - feminine "ženska) and paradigm that is list of endings. Theoretically names can have plural form so there are 14 ending for every paradigm (7 for singular and 7 for plural).

Paradigms table

Id	Mnem	Tag	Gend.	Paradigm (endings)
1.	Alex	x	m	0 a u a 0 u om i a ima e i ima ima
2.	Nik	N	M	0 a u a 0 u om ovi ova ovima ove

There are totally 46 paradigms comprising 29 male names paradigms, 11

male family names paradigm and 6 female names paradigm.

4. Morphological processing

The goal of automatic morphological processing is to perform automatically a morphological analysis and/or generation of some word-form. Moprhological analysis includes identifying the base word-form for a given word-form by recognizing some of its gramatical features and type of paradigm. Morphological generation include deriving all word-forms from a given basic form.

Example: applying paradigm endings to the noun (name) word-form generation

Name: Gubec
Paradigm: bec pca pcu pca pče pcu pcem pci baca pcima pce pci pcima pcima
Declension:

	singular	plural
N	Gubec	Gupci
G	Gupca	Gubaca
D	Gupcu	Gupcima
A	Gupca	Gupce
V	Gupče	Gupci
L	Gupcu	Gupcima
I	Gupcima	Gupcima

There are two steps included in word-form generation:

- firstly, for a given word, the nominativ ending is moved from the basic form (nominativ case) in order to determine begining of a word
- secondly, it is possible to generate all word-forms for a given begining of a word combining it with all coresponding endings including singular and plural form.

It should be emphasized that the beginings and endings are not equivalent to morphems.

5. Results and analysis

The morphological dictionary comprises xxxx names which produce over xxx wordforms. Linguistically, there are some problems causing bad results such as:

- the vocative problem,
- using alternative paradigm,
- source data errors.

The vocative case is the greatest problem in deciding to which paradigm a name belongs. In the Croatian language the probability of usage of the real vocative form instead of the nominative form in spoken language is very small, and in the written texts (except in school textbooks) is almost zero. Alternative paradigm was used only for alternative vocative forms and sometimes for the plural of family names. Because we couldn't get access to the social security data base or the data base of the ministery of interiors we were forced to use other publicly available sources which are much less accurate then the mentioned data bases.

6. Some Statisical Data

The most frequent female names paradigms

Paradigm	No.	No. of Occurences
a	5192	41569
s	359	8673
r	492	5259
n	1107	957
k	101	186

The most frequent family names paradigms

Paradigm	No	No. of Occurences
o	26790	82566

d	9654		n	2	
gg	4859				77759
x	2259				42457
bc	1320				31472
m	1440				20569
v	2940				17571
g	1599				14504
i	567				10993
c	455				9401
l	463				7778
k	822				6835
j	225				3940
ec	122				3830
ž	32				1988
y	211				615
h	116				595
vs	246				408
sc	9				318
šc	13				307
ža	10				232
sy	90				210
p	6				17
e	2				5
be	5				23

7. Conclusion

Until now there was no lexical (and inflectional) database for personal names, although there are several ordinary lexical and inflectional databases for the Croatian language. It should be emphasised that it is possible to use this database for a different purposes such as: additional source for Croatian spelling checker preparation, generative database for deriving all the possible forms of Croatian personal names as well as retrieving aid to search for all forms of a certain personal name i.e. as the personal name recognition module.

7. References

1. Boras, D., Kržak, M: Rječnička baza hrvatskog književnog jezika. Informatologia Yugoslavica, vol 17 (3-4), pp.223-242, 1985.
2. Kržak, M. Serbo-Croatian Morpho-Spelling. Computer Processing of Language Data (Proceedings of the IV-th Conference), pp. 207-214, Portorož, 1988.
3. Koskenniemi, Kimmo. Two-level morphology: A General Computation Model for Wod-Form Recognition and Production. Ph.D. thesis, University of Helsinky, Helsinky, 1983.
4. Lauc, D., Lauc, T., Boras, D., Ristov, S. : Developing Text Retrieval System using Robust Morphological Parsing. Proceedings of the 19th International Conference on Information Technology Interfaces, Pula, 1998.
5. Lopina, V. Dvorazinski model morfološkog opisa. Obrada jezika i prikaz znanja (zbornik, ur. Tkalac, Slavko – Tuđman, Miroslav), str. 75-80. Zagreb: Zavod za informacijske studije, 1993.

6. Thompson, P., Dozier, C. Name Recognition and Retrieval Performance. In Strzalkowski, T. ed., Natural Language Information Retrieval, pp. 261- 272, London: Kluwer Academic Publishers, 1999.