A computational morphology approach to Croatian noun inflection — the case of gender assignment

Milena Mihajlović

Wiener Linguistische Gazette
Institut für Sprachwissenschaft
Universität Wien
Special Issue 78A (2014): 120-127

Abstract

The primary aim of this paper is to present a research on noun inflection and gender in the Croatian language through the prism of computational morphology. The secondary goal of the paper is to set the basis for the computational model that would be able to predict gender. The research in question represents a network of different theoretical approaches, in particular Network Morphology, recent Natural Morphology account of Serbo-Croatian (Radisavljević, 2013), and the approach that served as a framework for the Croatian National Corpus.

1 Introduction

'Gender is arguably the most puzzling and fascinating of all grammatical categories'
(Corbett, 1991, p. 1)

Croatian is a South Slavic language. It is an inflecting-fusional language, meaning that the Croatian morphology is characterized by rich inflection, where one morpheme can denote different grammatical properties (e.g. the inflectional suffix *-a* in *rek-a* 'river' stands for three pieces of information: gender (feminine), case (nominative) and number (singular).

Clearly, the native speakers of the Croatian language acquired (easily) the knowledge on noun inflection and gender. In contrast, it is quite difficult to devise a computational model that would be able to predict both the inflection and the gender of nouns. Why is this so? The answer to this question will be provided in the next section.

Theoretical ground is based on the three frameworks: descriptive Croatian grammar (Barić et. al, 1997), Network Morphology and Natural Morphology. Traditional grammar offers us descriptions of the language features, but it is hardly used for the computational analysis because of the vagueness of the descriptions (see Tadić, 1994). Network Morphology is a computational approach to noun classification built primarily for the Russian language. Because of these two characteristics, being both a computational approach and devised for a

Slovene language, it is important for the research on the Croatian computational morphology. The importance of Natural Morphology for the research in question is twofold: first, it gives an insight into the unmarked (natural, cognitively easy) processes within the morphology, and second, the notion of productivity play a major role within this framework, and this will prove to be significant for the current research. In addition, the theoretical basis of the Croatian National Corpus is a computational model which is of importance for the current paper inasmuch as the focus of the research is on the Croatian nouns.

2 Data

In the Croatian language nouns have gender as an *inherent* property, and number and case as *morphological* properties. Moreover, Croatian has three genders – masculine, feminine and neuter grammatical gender, two values for the category of number – singular and plural, and seven cases in which nouns can vary (in singular, as well as in plural), see Table 1.

Sing.	*Class a*			*Class e*	*Class i*
	Masculine animate inanimate		Neuter		
N	Student 'student'	*Zakon* 'law'	sel-**o** 'village'	sten-**a** 'rock'	Mladost 'youth'
G	*student*-**a**	zakona-**a**	sel-**a**	sten-**e**	mladost-**i**
D	student-**u**	zakon-**u**	sel-**u**	sten-**i**	mladost-**i**
A	*student*-**a**	*Zakon*	sel-**o**	sten-**u**	Mladost
V	student-**u**	zakonu-**u**	sel-**o**	sten-**o**	mladost-**i**
L	student-**u**	zakon-**u**	sel-**u**	sten-**i**	mladost-**i**
I	student-**om**	zakon-**om**	sel-**om**	sten-**om**	mladost-**i**

Table 1: Noun inflection classification in Croatian – singular

Pl.	*Class a*			*Class e*	*Class i*
N	student-**i**	zakon-**i**	sel-**a**	sten-**e**	mladost-**i**
G	studenat-**a**	zakon-**a**	sel-**a**	sten-**a**	mladost-**i**
D	student-**ima**	zakon-**ima**	sel-**ima**	sten-**ama**	mladost-**ima**
A	student-**e**	zakon-**e**	sel-**a**	sten-**e**	mladost-**i**
V	student-**i**	zakon-**i**	sel-**a**	sten-**e**	mladost-**i**
L	student-**ima**	zakon-**u**	sel-**ima**	sten-**ama**	mladost-**ima**
I	student-**ima**	zakon-**om**	sel-**ima**	sten-**ama**	mladost-**ima**

Table 2: Noun inflection classification in Croatian – plural

Additionally, nouns are organized into three inflectional classes (cf. Tables 1 & 2) according to the traditional grammar (Barić et al., 1997). Namely, nouns are grouped into three declensional classes, named after the ending in genitive singular case, (e.g. G singular *igrač-a* 'player-g.sg' , *žen-e* 'woman-g.sg', *kost-i* 'bone-g.sg' – class a, class e, class i). Class *a* is the largest class and is further divided into two subclasses according to the gender of the nouns (masculine and neuter). Class *e* comprises mostly of feminine nouns ending in *-a*, while class *i* consists of feminine nouns ending in a consonant (Barić, 1997). Despite this, there are many subclasses that do not exactly fit the basic patterns of inflection, thus, have patterns of their own.

Now that the basic features of the Croatian language are explained, I will introduce some examples of gender assignment disparity which are especially challenging for the computer analysis. First, the cases of diminutives in *–če* are interesting, given that they all require neuter agreement, but not all of them refer to an inanimate object. For example, *momče* 'boy', is assigned neuter gender, nonetheless, it refers to a male person.

The noun *pijanica* 'drunkard' can be accompanied by both feminine and masculine adjectives, as it could refer both to a man and a woman. However, in either case, it inflects like most feminine nouns (belongs to the class *e*). Interestingly enough, feminine agreement does not necessarily mean that it refers to a female, sometimes it could also refer to a male.

Many languages do not distinguish gender in the plural (e.g. German). In contrast, Croatian belongs to the group of languages that do have it. In addition, in Croatian there are nouns that change their gender in the plural. For instance, the masculine noun 'kino' *cinema* changes its gender from masculine in the singular to neuter in the plural.

3 Classification of Serbo-Croatian nouns inflection according to the framework of Natural morphology

There are three subtheories within the framework of Natural Morphology: *universal markedness* (deals with system independent naturalness in morphology), *typological adequacy* (concerned with naturalness when speaking on the level of language typology) and *language specific system adequacy* (investigates morphological naturalness within a particular language system). Basically, sets of similar inflectional paradigms form inflectional classes in hierarchical order: macroclass, class, subclass, (subsubclass, if necessary, etc.), microclass. All classes are defined by implicational paradigm structure conditions (cf Wurzel 1984, Kilani-Schoch & Dressler 2005). 'The more genders a language has, the more macroclasses and productive microclasses it may have' (Dressler & Thornton, 1996 as cited in Radisavljević, 2013, p. 38).

Recently, a Master Thesis on Serbo-Croatian noun inflection is completed within the framework of Natural Morphology. It provided us with new insights into the morphology of the Serbo-Croatian. Namely, this classification is gender-based, with three macroclasses (including both productive and unproductive microclasses) being settled and defined. As already pointed out in the introductory section, the importance of Natural Morphology is twofold. Most importantly, it could have an impact on the design of the computational model that would be able to predict gender in Croatian since productive and unproductive classes could be dealt with in a different way, thus rendering the model simpler while at the same time making it more accurate.

4 Gender assignment

The purpose of this section is to present two computational morphology approaches to gender assignment – Network morphology and the approach implemented into the Croatian National Corpus.

4.1 Network morphology approach to gender assignment

The aim of this subsection is to illustrate how the Network morphology approach can be used to account for the noun inflection classification and gender assignment in Croatian. The basic concept underlining this theory is that a word inherits the properties from the node put in a higher place in the hierarchy, with only exceptional properties to be added further. If, however, something specified in the upper node does not hold for that particular node, the local information may override the inherited information (default). The approach itself is

based on the formal language DATR, which allows for computer interpretation. In short, the main constituents of Network Morphology include *values* (contain either atoms or list of a sequence of atoms, where atoms are considered to be undividable objects), *attributes* (might be either atoms or may consist of list of atoms), *facts* (consist of a pair of attribute and value), *nodes* (locations where facts are stored), and finally, *networks* which consist of the relationships between nodes and facts (Corbett & Fraser, 1993). Default inheritance is defined in Corbett & Fraser (1993: 120) in the following way:

> *If X and Y are nodes, X may inherit from Y if a fact identifying Y as an inheritance source is included at X. All attributes: value pairs at Y become available at X, except those having an attribute which is already present in an attribute: value pair at X.*

The notion of default could be illustrated with a well-known example involving the noun *penguin*. Penguin is a noun inherits all the properties of the noun *bird*, but would override the default that a bird *flies*, since it cannot fly (Corbett & Fraser, 1995). The main idea of Network morphology approach is thus related to inheritance. Certain defaults hold for the majority of examples. However, there is a possibility of defaults being overridden with the exceptions defined locally. Finally, gender is determined by either the semantics of the noun or its declensional class membership.

(1) žena (cf. Brown & Hippisley, 2012: 140):

<> == NOUN
<gloss> == woman
<root all> == žen
<sem sex> == female

The example above shows that semantic gender (*<sem sex> == female*) could be used for declensional class assignment.

(2) Student (cf. Brown & Hippsley, 2012: 53):

<> = NOUN
<declensional class> == N_I:<mor>
<gloss> == student
<root> == student

The example in (2) illustrates how gender could be inferred from declensional class: it belongs to class *I* (*<declensional class> == N_I:<mor>*), thus it has masculine gender.

However, an interesting example would be the noun *muškarč-in-a* (*-ina* is a typical augmentative suffix, see e.g. Barić et al., 1997) which denotes not only a man, but a manly man. It takes feminine agreement (and thus have feminine gender), but it denotes a male. Therefore, both the declensional class and the semantic gender have to be specified. It would be defined as:

(3) Muškarč-ina (cf. Brown & Hippisley, 2012: 141):

<> == NOUN

<gloss> == man

<declensional_class> == N_III : <mor>

<root all> == muškarčin

<sem sex> == male

Network morphology shows how the dynamics between gender and declensional class looks like in terms of computational linguistics.

4.2 Croatian National Corpus – the approach to gender assignment

While traditional grammar, as we have seen, uses the gender as its onset point in defining declensional classes, the defaults-based approach works in the opposite way – from a declensional class to gender assignment. The Croatian National Corpus could be regarded as a mixture of the traditional and default-based approaches. There is an exhaustive *list* of possible stems, *list* of endings and combination *rules*. Although the basic patterns for the inflection of nouns are taken from the traditional grammar, they are adapted for computer analysis (Tadić, 1994). Within this approach, the information about gender is listed only in case when it is not possible to predict it from the declensional class.

5 A computational model for gender inducement

The gender of nouns is to be induced from:

1. Proper noun endings:

The difference should be obtained between suffixed and unsuffixed nouns where e.g. both the unsuffixed noun *most* ‘bridge’ and the suffixed one *radost* ‘joy’ end in *–ost*. However, the unsuffixed word would take masculine agreement (except *kost* “bone” – takes feminine agreement), while the suffixed would take the feminine agreement. Furthermore, the difference between definite and indefinite adjectives should be accounted for. The feature that

is still notable on the morphological level in the first noun inflectional class (most nouns terminating in a consonant with masculine gender).

(4) lep most & lep-i most vs. lep-a rad-ost
'a beautiful bridge' 'the beautiful bridge' 'beautiful joy'

2. Agreement with adjectives:

Since gender is defined through the agreement with adjective words, it is thus *natural* to use agreement with attributive and predicative adjectives so as to induce gender. In this respect the Croatian National Corpus is to be accessed, not for the sake of analysis of the existing grammar rules but, to build a model that is going to illustrate what people actually use.

The former is thus related to the productivity as the basic concept of Natural Morphology (since productive and unproductive classes will have different treatment) as well as on the application of the default inheritance, similarly to what has been done within Network Morphology. The second part (agreement with adjectives) could also be assigned to Natural Morphology, given that fact that it has been proved to be a natural way for the gender acquiring.

6 Conclusion

To conclude, gender is viewed through the prism of different theoretical approaches, Natural Morphology, Network Morphology, and the approach used for Croatian National Corpus, all with the intention to reach the most effective way for gender assignment.

Finally, it is described how a computational model that will be able to predict gender of Croatian nouns is meant to be constructed. The implementation, however, is not yet fully conducted, which is the reason why the paper does not contain the details regarding the application of the model.

References

Argus, Reili (2007): *Acquisition of morphology in Estonian* (Doctoral dissertation, Tallinn University). http://e-ait.tlulib.ee/113/2/argus_reili2.pdf [January 2014]

Barić, Eugenija, Lončarić, Mijo, Malić, Dragica, Pavešić, Slavko, Peti, Mirko, Zečević, Vesna & Znika, Marija (1997): *Hrvatska gramatika*. Zagreb: Školska knjiga.

Brown, Dunstan & Hippisley, Andrew (2012): *Network morphology: a defaults-based theory of word structure*. Cambridge: Cambridge University Press.

Corbett, Greville (1991): *Gender*. Cambridge: Cambridge University Press.

Corbett, Greville & Fraser, Norman (1993): Network Morphology: a DATR account of Russian nominal inflection. *Journal of Linguistics, 29* (1), 113-142.

Corbett, Greville & Fraser, Norman (2000): Gender assignment: a typology and a model. In: Gunter Senft (Ed.), *Systems of nominal classification: Language, Culture and Cognition 4* Cambridge: Cambridge University Press 293-325.

Dressler, Wolfgang, Kilani-Schoch, Marianne, Gagarina, Natalia, Pestal, Lina & Pöchtrager, Markus (2006): On the Typology of Inflection Class Systems. *Folia Linguistica*, *40*, 51-74.

Dressler, Wolfgang (2006): Natural Morphology. In: Keith Brown (Ed.), *Encyclopedia of Language and Linguistics* (2nd ed. , Vol. 8). Amsterdam: Elsevier, 539-540.

Fraser, Norman & Corbett Greville (1995): Gender, animacy, and declensional class assignment: a unified account for Russian. In: Geert Booij & Joop van Marle (Eds.), *Yearbook of Morphology 1994*. Dordrecht: Kluwer, 123-150.

Radisavljević, Tijana (2013). *Productivity in Serbian Inflection and Derivation* (Master thesis, University of Vienna).

Tadić, Marko (1994): *Računalna obradba morfologije hrvatskoga književnoga jezika* (Doctoral dissertation, University of Zagreb). http://www.hnk.ffzg.hr/txts/mt-dr-le.pdf [January 2014].