Univerza v Ljubljani, skupni interdisciplinarni program druge stopnje Kognitivna znanost
v sodelovanju z Universität Wien, Sveučilište u Zagrebu, Univerzita Komenského v Bratislave in Eötvös Loránd Tudományegyetem

Milena Mihajlović

Računalniška morfologija in spol pri sklanjanju samostalnikov v hrvaškem jeziku

Magistrsko delo

Mentor:
Akad. Prof. Dr. Janez Orešnik

Somentor:
Emer. o. Univ.-Prof. Mag. Dr. Phil. Dr. h.c. Wolfgang U. Dressler

Ljubljana, 2015
Univerza v Ljubljani, skupni interdisciplinarni program druge stopnje Kognitivna znanost

v sodelovanju z Universität Wien, Sveučilište u Zagrebu,
Univerzita Komenského v Bratislave in Eötvös Loránd
Tudományegyetem

Milena Mihajlović

Računalniška morfologija in spol pri sklanjanju samostalnikov v hrvaškem jeziku

Magistrsko delo

Ljubljana, 2015
Abstract:

The aim of this thesis is to explore the Croatian noun inflection classes, with focus on the category of gender. Moreover, the aim is to build a computer model that would be able to induce gender. Since the inflection of nouns in this thesis tackles Computer Science, Linguistics and therefore Cognitive Science, a further aim is to explore cognitive correlates of the acquisition of inflectional morphology.

Keywords:

gender, noun inflection, computational model, cognition, croatian

Abstrakt:

Namen tega magistrskega dela je raziskati sklanjanje samostalnikov v hrvaškem jeziku s poudarkom na kategoriji spola in na podlagi tega narediti računalniški model, ki bo spol samostalnika samodejno induciral. Sklanjanje samostalnikov v tem magistrskem delu zajema jezikoslovje in računalništvo, ki sta del kognitivne znanosti. Nadaljnji cilj tega raziskovalnega področja je raziskati kognitivne korelate v zvezi z usvajanjem morfologije, še posebno spola.

Ključne besede:

spol, pregibanje, računalniška morfologija, kognicija, hrvaški
Acknowledgment

I wish to express an infinite gratitude to my supervisor and co-supervisor, Prof. Dr. Janez Orešnik and Emer. o. Univ.-Prof. Mag. Dr. Phil. Dr. h.c. Wolfgang U. Dressler for all the productive talks and each and every given piece of advice. Moreover, my gratitude to the Institute for Corpus Linguistics and Text Technology (ICLTT) for all the computational support is not smaller.

Furthermore, I wish to thank to my great professors in Ljubljana and my dear friends and colleagues Tijana Radisavljević, Nejc Trnjanin and Čedomir Oparnica for being willing to help and friendly advise whenever I was in need.
Contents

1. Introduction .................................................................................................................. 5
2. Approaches to Croatian noun inflection .................................................................. 6
   2.1. Data – descriptive linguistic approach ................................................................. 6
   2.2. A Natural Morphology account on Serbo-Croatian noun inflection classification ... 10
      2.2.1. The framework of Natural Morphology .......................................................... 10
      2.2.2. Productivity ................................................................................................. 12
3. Gender .......................................................................................................................... 18
   3.1. Defining gender in Croatian .................................................................................. 18
   3.2. Gender assignment in Croatian ............................................................................ 19
   3.3. Network morphology approach to gender assignment ......................................... 20
   3.4. Croatian National Corpus – approach to gender assignment ............................... 23
4. Model ............................................................................................................................ 24
   4.1. Background ........................................................................................................... 24
   4.2. Methods ............................................................................................................... 24
5. Croatian noun inflection with regard to the model .................................................. 26
   5.1. Proper vs. common nouns .................................................................................... 26
   5.2. Semantic level ...................................................................................................... 27
   5.3. Morphological level ............................................................................................. 28
   5.4. Phonological level ............................................................................................... 31
   5.5. Isolated groups ..................................................................................................... 32
   5.6. Regulary groups .................................................................................................. 32
   5.7. Agreement with adjectives .................................................................................. 33
6. Cognitive aspects of noun inflection in Croatian ................................................... 34
   6.1. Language acquisition .......................................................................................... 34
      6.1.1. Acquisition of inflectional morphology .......................................................... 35
      6.1.2. The acquisition of gender in Croatian ............................................................ 38
   6.2. Gender – arbitrary or semantically motivated? .................................................... 41
7. Conclusion .................................................................................................................... 43
8. Bibliography ................................................................................................................ 45
9. Appendix ...................................................................................................................... 49
   A. The Python code ..................................................................................................... 49
   B. The JavaScript code ............................................................................................... 55
1. Introduction

Croatian belongs to the South Slavic language subgroup. It is an inflection rich language, where one morpheme could stand for different grammatical properties (e.g. the inflectional suffix -a in stolić-a “chair” carries three pieces of information: feminine gender, nominative case and singular form).

The subject of our interest in this thesis will be the gender of nouns. It will be tackled through the prism of computational morphology. Moreover, linguistics and computer science are interwoven in cognitive science. Therefore, cognitive aspects of gender will be explored at the end of this thesis.

Gender of Croatian nouns is very interesting to explore. Traditionally, Croatian noun inflection system is seen as gender-based, i.e. gender is regarded as the main factor that governs the inflection of nouns. However, gender alone is not sufficient for the proper description of the noun inflection. Furthermore, a classification based exclusively on gender is problematic for computer analysis.

According to Hockett (1985 as cited in Corbett, 1991, p. 4) “genders are classes of nouns reflected in the behavior of associated words.” Here, “the behavior of associated words” refers to agreement, since gender of nouns govern the noun-adjective agreement, i.e. the associated adjective receives the same gender as the noun to which it refers, e.g. in the example (1), the adjective lep “beautiful” receives feminine gender, since the noun kuća “home” has feminine gender.

(1) lijep-a kuć-a
   beautiful-Fem.N.Sg house-Fem.N.Sg
   “beautiful house”

On the basis of the afore-stated, the model is meant to be built in such a way that gender is induced from the agreement with adjectives and from the rules which are going to be formulated on the basis of the traditional Croatian grammar, Network morphology and Natural Morphology, along with the Croatian National Corpus (in further text – Corpus) (Tadić, 2009). Speaking of rules, it is important to stress that in this thesis rules are understood as patterns that differ with regard to the productivity level (see Dressler & Laaha, 2012).

First, I will present basic information concerning morphology and noun inflection classification of the Croatian language. Second, I will present Natural Morphology approach to Serbo-Croatian noun inflection. It aims for the natural morphological patterns, thus offering a distinctive view of the noun inflection classification in Croatian. Only recently we were provided with this account (MA thesis written by Radisavljević, 2013). The framework of Natural Morphology has a twofold significance. Namely, not only does it offer a different and renovated version of noun classification, but it also recognizes productivity as an important governing principle in morphology.

Afterwards, the focus will be shifted to gender assignment in Croatian. In this respect, the framework of Network morphology will be presented in addition to the one utilized by the
Corpus. The former is expected to provide novel insights into how the language *breathes* on the level of morphology. *Default inheritance* plays a major role in the Network morphology approach and it “allow[s] a generalization to hold for most cases rather than all cases” (Brown & Hippisley, 2012, p. 33). Namely, Croatian noun inflection is closely related to phonology, which allows for generating of inflectional forms on the basis of gender and/or phonology of nouns with the help of default rules. Moreover, a model for the Croatian gender inducement will be introduced and discussed.

Many languages do not have gender distinctions in plural, e.g. German. On the other hand, Croatian belongs to the group of languages that do have it. Using it as an example, my final goal will be to investigate the cognitive implications of gender and to explore the acquisition of gender in Croatian.

2. **Approaches to Croatian noun inflection**

While true that there is a tight connection between noun classification and gender, gender alone is not enough for noun classification. Since gender is highly incorporated in the noun inflection system, this section will provide the account on the basic structure of the Croatian noun inflection system. In order to gain more knowledge on noun inflection in Croatian, the traditional descriptive linguistic approach and the Natural Morphology approach to Croatian noun inflection classification will be presented. The reasons the Natural Morphology approach is chosen will be explained below (section 2.2).

2.1. **Data – descriptive linguistic approach**

In Croatian, gender is an inherent property of nouns, while number and case are morphological properties. This means that a noun in Croatian can *belong to* masculine, feminine or neuter gender and that the very same noun can usually be expressed both in singular and in plural within seven possible cases, see Table 1. Moreover, there are three main inflectional classes nouns can belong to (cf. Table 1) according to Barić et al. (1997). Additionally, there are many subclasses that do not exactly fit the basic patterns of inflection and have patterns on their own. As defined by Aronoff (1994, p. 64) “an inflectional class is a set of lexemes whose members each select the same set of inflectional realization.” Those three declension classes are named after the ending vowel in the genitive singular, (e.g. G singular *pjevač*–a “singer,” *žen*–e “woman,” *mladost*–i “youth” – *class a, class e, class i*). Although, almost every loanword is fully adapted to one of the existing morphological patterns of declension, it is important to point out that there is a certain amount of indeclinable nouns (e.g. personal foreign names such as *Ingrid, Karmen, Nives*) (Barić et al., 1997).

Class *a* is divided into two subclasses according to the gender – masculine and neuter. Class *e* consists mostly of feminine nouns ending in -a, while class *i* comprises of feminine nouns ending in a consonant (Barić, et al., 1997).

---

1 It is important to note that some grammars recognize four different inflectional classes. Basically, some put the two different subclasses of the class *a* into two distinct classes (cf. e.g. Stevanović, 1970).
Also, it is vital to point out that there is the problem of animacy in the Croatian language. Namely, semantic differences between animate and inanimate objects is highly important for the inflection of nouns in Croatian, since the accusative singular is formed in a different way for animate and inanimate masculine nouns: it is identical to the genitive singular for the former nouns, and to the nominative singular for the latter ones. Corbett (1991) regards this property of masculine gender as having two subgenders, since those subclasses of nouns also take different agreement, just as shown in Table 1 (with the animate noun student “student,” and the inanimate noun zakon “law”). Corbett defines subgenders as “agreement classes which control minimally different sets of agreement, that is, agreements differing for at most a small proportion of the morphosyntactic forms of any of the agreement targets” (Corbett, 1991, p. 163). In the plural, Croatian does not make a difference in terms of animacy.

In line with what has already been specified above, most nouns are assigned to gender by semantic rules – males and females are assigned masculine and feminine gender respectively, while the gender of the other nouns is assigned according to phonological and morphological rules. Therefore, animacy appears to be of great importance for noun inflection. Moreover, it is a challenge for computer analysis since animacy cannot be inferred in any other way except lexically.

Traditional grammar is usually regarded as gender-based, as gender plays an important role in noun classification, since it is the only property nouns inherently possess. Moreover, most masculine nouns end in consonant, most neuter nouns in –o/-e, while most feminine nouns end in –a (Barić et al., 1997). As already mentioned above, gender is also of great importance for the agreement and vice versa, e.g. moj slug-a “my servant,” the noun “servant” requires a masculine pronoun, although it inflects like most feminine nouns (class e).
Since gender is defined through agreement it should be demonstrated how the relationship between gender and agreement looks like. The examples below (from 2 through 4) illustrate the relationship between gender and agreement, where grad “town” belongs to the masculine, selo “village” to the neuter and soba “room” to the feminine gender. In all cases the adjective agrees in number, case and gender with the noun, and here the adjective is a form of the possessive pronoun moj “my.”

(2) moj grad “my town”

<table>
<thead>
<tr>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>N moj grad</td>
<td>moji grad-ov-i</td>
</tr>
<tr>
<td>G mog grad-a</td>
<td>mojih grad-ov-a</td>
</tr>
<tr>
<td>D mom grad-u</td>
<td>mojim grad-ov-ima</td>
</tr>
<tr>
<td>A moj grad</td>
<td>moje grad-ov-e</td>
</tr>
<tr>
<td>V moj grad-u</td>
<td>moji grad-ov-i</td>
</tr>
<tr>
<td>L mom grad-u</td>
<td>mojim grad-ov-ima</td>
</tr>
<tr>
<td>I mojim grad-om</td>
<td>mojim grad-ov-ima</td>
</tr>
</tbody>
</table>

(3) moja soba “my room”

<table>
<thead>
<tr>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>N moja sob-a</td>
<td>moje sob-e</td>
</tr>
<tr>
<td>G moje sob-e</td>
<td>mojih sob-a</td>
</tr>
<tr>
<td>D mojoj sob-i</td>
<td>mojim sob-ama</td>
</tr>
<tr>
<td>A moju sob-u</td>
<td>moje sob-e</td>
</tr>
<tr>
<td>V moja sob-o</td>
<td>moje sob-e</td>
</tr>
<tr>
<td>L mojoj sob-i</td>
<td>mojim sob-ama</td>
</tr>
<tr>
<td>I mojom sob-om</td>
<td>mojim sob-ama</td>
</tr>
</tbody>
</table>

(4) moje selo “my village”

<table>
<thead>
<tr>
<th>Singular</th>
<th>Plural</th>
</tr>
</thead>
<tbody>
<tr>
<td>N moje sel-o</td>
<td>moja sel-a</td>
</tr>
<tr>
<td>G mog sel-a</td>
<td>mojih sel-a</td>
</tr>
<tr>
<td>D mom sel-u</td>
<td>mojim sel-ima</td>
</tr>
<tr>
<td>A moje sel-o</td>
<td>moja sel-a</td>
</tr>
<tr>
<td>V moje sel-o</td>
<td>moja sel-a</td>
</tr>
<tr>
<td>L mom sel-u</td>
<td>mojim sel-ima</td>
</tr>
<tr>
<td>I mojim sel-om</td>
<td>mojim sel-om</td>
</tr>
</tbody>
</table>

What I will illustrate with the following examples are gender discrepancies that the Croatian noun system exhibits, and which are especially relevant for this thesis since they represent a
challenge for a computer analysis, and are at the same time interesting to be explored in terms of cognitive science.

(5) *ovaj slug-a*
    this.Masc.N.Sg servant-Masc.N.Sg
    “this servant”

In the example (5) *slug-a “servant”* is masculine, although the adjective pronoun *ovaj “this”* has masculine gender, the noun *sluga “servant”* inflects like a feminine noun (see example 3). Therefore, it is important for this noun to be marked as having masculine gender in the noun inflection model.

Another interesting example is in (6). The neuter noun *momče “boy”* inflects like a neuter noun (see example 4), although it denotes a *male* person.

(6) *ov-o momč-e*
    this-Neut.N.Sg boy-Neut.N.Sg
    “this boy”

The example in (7) is interesting since the noun *pijanica “drunkard”* like e.g. *varalica “cheater,”* can be accompanied both by feminine and masculine adjectives, as it could be related to a man or a woman (although it is interesting that even when there is a feminine adjective it does not necessarily refer to a feminine person), but with respect to inflection behaves like most feminine nouns ending in *-a* (class e). Stevanović (1970) wrote that when such a noun refers to a male it could take both feminine and masculine agreement, however when it refers to a female it could take only feminine agreement. Still, Barić et al. (1997) noted that with the upper mentioned nouns it is more common to have feminine adjectives in both singular and plural whatever the gender of a noun is, even though both are possible.

(7) *ovaj pijanic-a*    *ova pijanic-a*
    this.Masc.N.Sg drunkard.Masc.N.Sg    this.Fem.N.Sg drunkard
    “this drunk man”    “this drunk woman/man”

As Anderson (1992 as cited in Corbett, 2006) puts it, agreement is “a quite intuitive notion which is nonetheless surprisingly difficult to delimit with precision” (p. 4).

In addition, there are nouns that change their gender in the plural. The example in (8) illustrates that the noun *kino “cinema”* changes its gender from masculine in the singular to neuter in the plural.

(8) *ovaj kin-o*    *ova kin-a*
    this.Masc.N.Sg cinema-Masc.N.Sg    these.Neut.N.Pl cinema-Neut.N.Pl
    “this cinema”    “these cinemas”
2.2. A Natural Morphology account on Serbo-Croatian noun inflection classification

The purpose of this section is to describe a Natural Morphology approach to Serbo-Croatian noun inflection classification, which strives for the natural features in morphology, and is therefore, of great importance to both cognitive science and to this thesis. Moreover, it is of even greater importance to this thesis given the fact that productivity will play a significant role in the model being developed here, and that the productivity plays one of the most important roles in Natural Morphology (see the next section). In order to understand it, some basic information about the framework of Natural Morphology will be presented.

2.2.1. The framework of Natural Morphology

Sets of similar inflectional paradigms form inflectional classes (in the generic sense), in hierarchical order: macroclass, class (in the specific sense: similar to the traditional term of, e.g., the five Latin declension classes, where not all nouns of one class inflect in exactly the same way), subclass, (subsubclass, if necessary, etc.), microclass. All classes are defined by implicational paradigm structure conditions (Dressler et al., 2006, p. 52).

The concept of naturalness could be perceived on the levels of the three existing subtheories within this framework – universal markedness, typological adequacy and language specific system adequacy, which will be presented here briefly (cf. Dressler et al, 1987).

a. First subtheory – Universal markedness

The first subtheory was proposed by Willi Mayerthaler in 1981 and it deals with system independent naturalness in morphology (cf. Dressler, 2006). It is a preference theory, meaning that it settles what is natural within naturalness parameters. It is important not to understand those preferences as global preferences, but as the preferences for a given parameter (Dressler, 2006). Even though those preferences are found cross-linguistically, there is a possibility for a language to be found that is not in agreement with some of the preferences being natural for the majority of languages (see e.g. Radisavljević, 2013).

Here, I will present some of the most important parameters which are iconicity, morphosemantic transparency, morphotactic transparency, binarity and the parameter of optimal shape.

Iconicity: morphosemantically marked category has a preference to have a longer form compared to the unmarked one (Dressler, 2006). Plural form composed by suffixing, e.g. in Croatian grad

---

2 “Serbo-Croatian” or “Serbo-Croat” is a term for the languages of former Yugoslavia: Serbian, Bosnian, Montenegrin and Croatian. As the division between Serbian and Croatian is in morphology more noticeable on the level of verbs than on the level of nouns, in further work I will use the term “Croatian” for my own work, since I use the Croatian National Corpus. However, Radisavljević (2013) defined it as “Serbo-Croatian,” and it will be referenced as such here as well. Moreover, there are dictionaries that were written in the time when “Serbo-Croatian” was a politically correct title of the unified language, and they will also be referenced with such a name.
“city,” plural grad-ovi, is more iconic compared to the one formed adding a suppletive form, e.g. neut. sg. dijet-e “child,” fem. pl. djej-a. There are five basic techniques according to the preference of iconicy (Manova & Dressler, 2001):

**Addition:**

(9) masc. učitelj “teacher” → fem. učitelj-ica

**Substation:** e.g. in forming masculine and feminine forms of nouns:

(10) masc. neznan-ac “unknown male person” → fem. neznan-ka “unknown female person”

**Modification:**

(11) verb plesti “knit” → noun masc. plot

**Conversion:**

(12) adjective acc. neut. dobro “good” → noun neut. dobro “estate”

**Subtraction:**

(13) masc. sg. zemljanim “earthling” → masc. pl. zemljan

Morphosemantic transparency is another important parameter: complete transparency reflects fully transparent meaning of compositional forms. For example, the meaning of grad-ovi “cities” is fully transparent combination of the noun grad “city” and the suffix for making a plural form.

The parameter of morphotactic transparency: the more transparent the decomposability of forms is in terms of their opacity, the more it is natural. For instance, an adjective studentski “student” is more natural compared to the adjective knjiški “bookish” (student + ski -> studentski vs. knjiga + ski -> knjiški).

As for the parameter of binarity, it is stated that binary forms are more natural than tertiary or n-ary. For instance, half-compounds such as klima-uredaj “air conditioner” are often found in Croatian. There is a very limited and restricted use of those tertiary, e.g. crveno-bijelo-plava zastava “red-white-blue flag.”

Naturalness within the parameter of optimal shape: an optimal shape of a morpheme is considered to be one syllable, while for a word an optimal shape is one foot (Dressler, 2006). For instance, in Croatian affixes typically consist of one syllable, e.g. –ac, -će, -je, etc.

**a. Second subtheory – Typological adequacy**

The second subtheory is established by Wolfgang U. Dressler in 1985, and it deals with naturalness on the level of language typology (cf. e.g. Dressler, 2006). Hence, some morphological properties may be more natural for inflecting-fusional languages (Croatian), while others may be more natural for agglutinative languages. In this respect, with reference to data obtained by the Natural Morphology approach, strongly inflecting languages commonly have more macroclasses and more productive microclasses and thus more complex system on all levels. In the ideal inflecting-fusional type there are more productive inflectional categories and
rules, but more unproductive microclasses. At the opposite end would be an ideal agglutinating language, where all the morphological patterns are productive (Dressler et al., 2006). However, it is also truth what Dressler and Thornton (1996) stated: “The more genders a language has, the more macroclasses and productive microclasses it may have” (p. 23).

b. Third subtheory – Language specific system adequacy

*The structure of inflection class systems is a major object of the third subtheory of Natural Morphology.* (Dressler et al., 2006, p. 51)

The third subtheory was proposed by Wolfgang U. Wurzel in 1984, and its aim is morphological naturalness within a particular language system (Dressler, 2006). The major three concepts for defining naturalness on this level are defaults, morphological classes and productivity.

These concepts are of great importance for this thesis, as we are dealing with a language specific system of nominal inflection – Croatian. Forasmuch as Mayerthaler (1981 as cited in Radisavljić, 2013) “defined the productivity as a result of the naturalness phenomenon” (p. 11), and since productivity plays an important role for the model that will be presented in this thesis, I will devote a subsection to the notion of productivity (2.2.2.).

Recently, a Master’s thesis was written on Serbo-Croatian noun inflection within the framework of Natural Morphology by Radisavljić (2013), which provides us with new pieces of information on the Serbo-Croatian noun classification. Namely, Radisavljić (2013) proposed a new gender-based classification with three macroclasses with both productive and unproductive microclasses being settled and defined. A closer look into this account and its contribution to the model for predicting gender in Croatian will be provided below alongside the model (section 5).

Finally, I find the viewpoint of Natural Morphology important for the model design and the interpretation of the data, since productive and unproductive classes could be dealt in different ways. Therefore the model for predicting gender could be simpler and more accurate at the same time.

2.2.2. Productivity

The notion of productivity is one of the core concepts within the third subtheory of Natural Morphology (the theory of language-specific system). Along with the default and morphological classes, as Dressler put it (2003, p. 31): “… productivity should be taken as a constitutive primitive property of inflectional patterns (or rules or processes), in the same way as in the other components (or (sub) modules) of grammar.”

In this thesis, productivity is understood to be a gradual process that can be measured (Dressler, 2003). Measurement criteria will be provided below in section 2.2.2.1.

Productivity belongs to the potential system of grammar and in terms of Natural Morphology it is a property of *dynamic morphology*. It is the level of potential words, with productive and unproductive rules, categories and classes. Opposed and complementary to dynamic morphology stands *static morphology*, which is the level of already existing words in a lexicon. The linkage between static and dynamic morphology is achieved indirectly through productivity, together
with the concepts of type and token frequency. Namely, the more productive a morphological rule is the more words it will form, thus it leads to higher token frequency. On the other hand, high token frequency leads to decreasing productivity. This is because words built according to the productive rules at some point reach high frequency and become lexicalized and language learners memorize them as a whole. Therefore, the correlation between dynamic and static morphology vouch for a language change. As a matter of fact, once most productive rules become the unproductive ones.

Given the fact that in psycholinguistic models (those shaped according to generative grammar), productivity, although viewed as being quite a relevant property of morphological rules “has been subordinated to concepts of regularity and/or default” (Dressler, 2003, p. 32). Moreover, it seems relevant to stress the difference between productivity and regularity, as well as the difference between productivity and defaults.

*Regularity* must not be confused with productivity, since both productive and unproductive rules can have an output that is regular. Hence, the term “regularity” could be conceived as a hypernym of productivity (Dressler, 2003). Regularity could be understood in terms of homogeneity, in such a sense that decreasing the homogeneity leads to decreasing the regularity. For instance, Dressler (2003, p. 36) gives an example of Breton alternations, where “lenition of /p, t, k/ to /b, d, g/ is more regular than lenition of /b, d, g/ to /v, z/ and zero, respectively.”

Additionally, a difference must be made between *productive* and *default* rules (cf. Dressler, 2003; Radisavljević, 2013). A morphological rule could be a default rule (meaning that it is applicable to most words in a language), although it does not have to be a productive one. For instance, in Serbo-Croatian both neuter and masculine nouns could end in –o, and most of nouns that end in –o are neuter and it is a default neuter ending. However, only masculine in –o are *productive*. This means that if a loanword in –o would have to be accepted in Serbo-Croatian, it would take masculine agreement. Examples in (14) and (15) exemplify this claim on the level of loanwords.

(14) sako “jacket,” a noun that probably came into Serbo-Croatian from German (*der/das Sakko “sport coat,” from ital. *sacco “sack”*), and it belongs to the masculine gender (Vujaklija, 1991).

(15) bingo “bingo, jackpot” is masculine in Serbo-Croatian came into this language most probably from English, a genderless language, which is why this example is even more interesting.

Finally, the great importance of productivity is best illustrated with the example of the noun *auto “car,”* loaned from German. Since it ends in –o it should be assigned neuter gender by default. However, even though –o is a default ending for neuters in Croatian, and it belongs to the neuter gender in German (the source language) it was assigned masculine gender in Croatian. In addition, in Polish, the same noun *auto “car”* was assigned neuter gender. The reason why the same noun is neuter in Polish and masculine in Croatian is due to the fact that neuter microclass in –o is not productive in Croatian, while it is productive in Polish.
2.2.2.1. Measurement of word productivity

\textit{a. The first criterion – adapting a word with unfitting properties}

The first (a) is the most important criterion, since “a rule must have maximum productivity in order to overcome the two obstacles of foreignness and unfitting properties” (Dressler, 2003, p. 37). It is, as stated in Dressler (2003), Wurzel’s (1984) secondary productivity rule, and although Wurzel (1984) wrote only about the adaptation of unfitting phonological shapes, an example of which would be loanwords with ending vowel adapted in a target language, Dressler (2003) suggests that of equal importance is fitting of other criteria, particularly gender, with maintaining a word form or with a minimal adaptation of a form (either phonological or graphic). Examples in (16) and (17) illustrate phonological fitting.

(16) fem. \textit{bin-a} “stage” from German fem. \textit{Bühne} “stage” (Vujaklija, 1991)

(17) fem. \textit{vag-a} “weighing scale” from German \textit{Waage} “weighing scale” (Vujaklija, 1991)

Moreover, fitting is an indication of which inflectional classes are productive (Dressler 2003; Radisavljević, 2013). Croatian gender is an internal property of a noun, and therefore words loaned from a genderless language will be assigned masculine, feminine or neuter gender. Assigning gender to a noun from a genderless language is rightfully regarded as one of the most significant evidence for productive morphological rules.

The following examples illustrate gender fitting in Croatian:

If a novel noun ends in \textit{–a} it will be assigned to the third class (the default class for feminine nouns, see Table 1), and it will be assigned feminine gender, given that this is the only productive class for feminine (cf. Radisavljević, 2013), which is shown in (18) and (19).

(18) fem. \textit{džungl-a}, “jungle,” borrowed from English (genderless) (Vujaklija, 1991)

\begin{verbatim}
gust-a džungl-a
dense-Fem.N.Sg jungle-Fem.N.Sg
“dense jungle”
\end{verbatim}


\begin{verbatim}
kratk-a tank-a
short-Fem.N.Sg tanka-Fem.N.Sg
“short tanaka”
\end{verbatim}

If a noun terminates in \textit{–o/e/u/i}, masculine gender is assigned:

(20) masc. \textit{derb-i} “derby,” borrowed from English (Vujaklija, 1991):

\begin{verbatim}
odličan derb-i
great.Masc.N.Sg derby-Masc.N.Sg
“great derby”
\end{verbatim}
(21) masc. šo-u “show,” borrowed from English (Vujaklija, 1991):

\[ odličan \ šo-u \]
\[ great.Masc.N.Sg \ show- \ Masc.N.Sg \]
\[ v “great show” \]

(22) masc. kimono “kimono,” borrowed from Japanese (Vujaklija, 1991):

\[ šaren-i \ kimon-o \]
\[ colorful-Masc.N.Sg.\Def \kimono-Masc.N.Sg \]
\[ “the colorful kimono” \]

If a loanword ends in a consonant, it will be assigned masculine gender:

(23) masc. kardigan “cardigan,” borrowed from English.

\[ zeleni \ kardigan \]
\[ green.Masc.N.Sg.\Def \cardigan.Masc.N.Sg \]
\[ “the green cardigan” \]

**b. The second criterion – adapting a word with fitting properties**

The second criterion, represented by Wurzel’s (1984) primary productivity rule, is integration with fitting properties. Therefore, “integration must overcome only the obstacle of foreignness” (Dressler, 2003, p. 39).

As Radisavljević (2013) noted in her master thesis, most of those loanwords with fitting properties come from Italian and Spanish. Examples in (24) and (25) illustrate fitting of already existing properties in the target language.

(24) fem. štafet-a “relay” from Italian fem. staffetta (Vujaklija, 1991)

(25) masc. kabinet “office” from French masc. cabinet (Vujaklija, 1991)

Especially are interesting cases in which a loanword seems to have all the properties that fit to the target language system, but it is not integrated as such. For example, it has gender in the source language and it ends in a vowel or a consonant that is common in the target language. However, in the target language it is assigned different gender. One of such loanwords in Croatian is auto “car,” already discussed in section 2.2.2. (see example 15). This is an interesting noun since it ends in \(-o\), and even though \(-o\) is a default ending for neuters in Croatian, and it is neuter in German (the source language) it was assigned masculine gender. Moreover, in Polish, the same noun auto “car” was assigned neuter gender. Therefore, neuter microclass in \(-o\) is not productive (cf. Radisavljević, 2013).

(26) zelen-i aut-o

\[ green-Masc.N.Sg.\Def \textit{car}-Masculine.N.Sg \]
\[ “the green car” \]
c. The third criterion – inflection of indigenous neologisms

The third productivity criterion is the inflection of indigenous neologisms, i.e. conversion and abbreviations. Conversions, already mentioned in the section on the Natural Morphology framework (first subtheory of Natural Morphology, 2.2.1. a.), are presumed to be less natural than affixation by the universal naturalness preference parameter of iconicity. Therefore, conversions are less integrated into inflectional patterns. Substantivization (conversion of adjectives or other word-classes into nouns) and adjectivization (conversion of participles into adjectives) are typical types of syntactic conversion in Slavic languages (Radisavljević, 2013).

The following examples in (27) illustrate the conversion of adjectives into nouns. They keep the neuter gender but the word class is different.

(27) a) noun dobro-“goods” from adj. dobar, neut. dobro-“good”

<table>
<thead>
<tr>
<th>dobro- del-o</th>
<th>poljoprivredn-o dobro-o</th>
</tr>
</thead>
<tbody>
<tr>
<td>good-Neut.N.Sg deed-Neut.N.Sg</td>
<td>agricultural-Neut.N.Sg property-Neut.N.Sg</td>
</tr>
<tr>
<td>“good deed”</td>
<td>“agricultural property”</td>
</tr>
</tbody>
</table>

b) noun slatko-“marmalade” from adj. sladak “sweet,” neut. slatko-

<table>
<thead>
<tr>
<th>slatko jel-o</th>
<th>ukusn-o slatko-</th>
</tr>
</thead>
<tbody>
<tr>
<td>delicious-Neut.N.Sg marmalade-Neut.N.Sg</td>
<td>sweet-Neut.N.Sg dish-Neut.N.Sg</td>
</tr>
<tr>
<td>“delicious marmelade”</td>
<td>“sweet dish”</td>
</tr>
</tbody>
</table>

The next example shows the conversion of the present adverbial participle svetleći “illuminating” into an adjective.

(28) svetleći bilbord

<table>
<thead>
<tr>
<th>bilbord</th>
<th>svetleći bilbord</th>
</tr>
</thead>
</table>

Abbreviations are relevant since they could vouch for productive rules and classes in a language. The following examples (28-31) are taken from Radisavljević (2013, p.47):

(29) masc. autobus “bus,” abbreviated: masc. bus

(30) masc. fakultet/faksimil “faculty/facsimile,” abbreviated: masc. faks

(31) masc. profesor “professor,” abbreviated: masc. profa

(32) fem. diskoteka, “disco,” abbreviated: masc. disko, also form masc. diskać
d. The forth criterion – class shift of a paradigm

The fourth, hierarchically lower productivity criterion is class shift of a paradigm, typically from a less productive class to a more stable and a more productive microclass (Dressler, 2003). For instance, Radisavljević (2013, p. 48) notes that in one of the dialects in the South-Central Serbia mesto “place” is inflected as a feminine and gives an example used in ordinary speech: *na dv-e mest-e instead of the regular form na dv-a mest-a “on two places.”

Radisavljević (2013) illustrates class shift within the standard dialect as well. Such an example is given in (33). There happened a shift from the unproductive microclass of plurale tantum to the productive class e (see Table 1).

(33) regular form: acc. pl. neut. script-a “notes” (plurale tantum)
    often used in ordinary speech: acc. sg. fem. script-u.

e. The fifth criterion – word-formation

The last and hierarchically lowest criterion is the productivity of affixations, thus word-formation criterion. Although it is about word-formation and not directly on inflection, this criterion gives an account of productive inflectional microclasses (cf. Dressler, 2003).

Radisavljević (2013, p. 49) states that “Serbian ‘motion suffixes’ (or gender shifts) are connected to the two most productive classes, i.e. –ica, -ka, -kinja (class II.1) for feminine and –ac (class I.1) for masculine.”

(34) masc. lav “lion”
    fem. lav-ica “lioness”

(35) fem. lis-ica “fox”
    masc. lis-ac “male fox”

Productive suffixes have a syntactic function in adapting loanwords to the language specific syntactic patterns for number, case, tense, etc. As Dressler (2003) put it, “if a rule is not productive enough for fulfilling this function, then either a more productive rule takes over, or the new word remains uninflected, which may be, first of all, awkward for syntax” (p. 45). The consequence of this urge or tendency to have all the words inflected is putting the novel ones into the most productive inflectional patterns. This process was confirmed by the examples given in this section.

3 Class e and class a, respectively.
3. Gender

*Gender is arguably the most puzzling and fascinating of all grammatical categories.*

(Corbett, 1991, p. 1)

The aim of this chapter is to present the problem of defining gender in the Croatian language. Moreover, I intend to present two approaches to gender assignment. Namely, the first one is constructed within the Network morphology framework and the other is implemented within the Corpus.

According to Nichols’ (1992 as cited in Franceschina, 2005, p. 85), who surveyed 174 languages, “just over one-fourth of them had gender or some other form of nominal classification system.” However, although gender is mostly related to nominal system, it could as well be related to pronominal system. Furthermore, languages with only pronominal gender systems are seldom, and there are only five found in the Indo-European language family: Afrikaans, English, Manx, Persian and Yazgulyam (Audring, 2009). Implication: if a language has a gender distinction in nouns, it has it also in third singular pronouns, but not vice versa.

Gender is an intriguing feature as it is neither a totally formal, nor a totally semantic category of nouns (in most languages). “While there are no systems in which formal information alone is sufficient, we do find systems in which semantic information is sufficient for gender assignment, and it is these that are known as *semantic systems*” (Corbett & Fraser, 2000, p. 294). In addition, the Dravidian language Tamil is an example of a semantic gender assignment system.

3.1. Defining gender in Croatian

A place where many different definitions of gender could be found is literature. In Franceschina (2005) four different groups of gender definitions are listed with respect to different foci (e.g. the focus may be on gender manifestation or on the function of gender). However, as already defined in the introductory section, the broadly accepted definition of gender is the one by Hockett (1985 as cited in Corbett, 1991), where “genders are classes of nouns reflected in the behavior of associated words” (p. 4). Therefore, this definition brings out the importance of adjectives for the assignment of gender.

Analogous to what has been already shown above, there are three genders nouns could be assigned to in the Croatian language: masculine, feminine or neuter, both in the singular and in the plural.

There is much inconsistence in terms of defining gender in Croatian and Serbo-Croatian grammar books (cf. Nikolić, 2000-2003). First, there is a controversy around the question of so-called natural and grammatical gender. Namely, although a so-called *natural* gender is a lexico-semantical category, and not morphological, when it comes to concordance it might turn out to be important for the morphology (Nikolić, 2000-2003). There is no consensus about what exactly *natural* gender is. According to Nikolić (2000-2003), while some grammarians deem that there are two natural genders – masculine and feminine (e.g. Mrazović & Vukadinović, 1990 in Nikolić, 2000-2003) – corresponding to biological sexes, others deem that there are three (e.g.

Second, there is a disagreement in terms of criteria that should be applied when assigning gender. Authors of various Croatian dictionaries and grammar books define and assign gender differently. Namely, there are two criteria for the gender assignment: it is either with regard to the agreement with adjectives or with regard to their declension (Pišković, 2011). When it comes to a discrepancy between the two criteria, grammarians give precedence to one or the other criterion. For instance, if we take vojvoda “duke,” it would be assigned masculine gender according to the first criterion (takes masculine agreement), but feminine according to the second (it inflects like most feminine nouns ending in –a, class e).

(36) vojvoda “duke” žena “woman”

<table>
<thead>
<tr>
<th>Case</th>
<th>Noun Form</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>vojvod-a</td>
<td>žen-a</td>
</tr>
<tr>
<td>G</td>
<td>vojvod-e</td>
<td>žen-e</td>
</tr>
<tr>
<td>D</td>
<td>vojvod-i</td>
<td>žen-i</td>
</tr>
<tr>
<td>A</td>
<td>vojvod-u</td>
<td>žen-u</td>
</tr>
<tr>
<td>V</td>
<td>vojvod-o</td>
<td>žen-o</td>
</tr>
<tr>
<td>L</td>
<td>vojvod-i</td>
<td>žen-i</td>
</tr>
<tr>
<td>I</td>
<td>vojvod-om</td>
<td>žen-om</td>
</tr>
</tbody>
</table>

(37) odličan vojvod-a
excellent.Masc.N.Sg.Indef duke-Fem.N.Sg
’an excellent duke’

There is a small number of nouns such as sluga “servant” that impose a real problem to the gender assignment, since grammarians do not even agree on whether this noun refers both to males and females (e.g. Stevanović, 1970) or only to males (e.g. Simeon, 1969).

3.2. Gender assignment in Croatian

Corbett (1991) proposes two types of gender assignment: formal and semantic, where formal could refer to either phonological or morphological assignment (Corbett, 1991; cf. Manova & Dressler, 2001).

The following is an implementation of the upper-stated gender assignment rules:

Phonological assignment

a) Nouns terminating in a consonant are masculine
b) Nouns terminating in –a are feminine
c) Nouns terminating in –o/-e are neuter

---

4 This grammar book is the most used in secondary schools and at the Universities in Serbia.
If we apply those rules, we would probably be wrong with many Croatian nouns, since it is true that there are also a great number of masculine nouns in –o. Even though there are more nouns in –o/-e that are neuter, only masculine are productive (cf. Radisavljević, 2013). Phonological assignment is always a default assignment, and as for the defaults, it should be noted that there are strong and weak defaults.

**Semantic assignment**

a) Nouns denoting females are feminine  
b) Nouns denoting males are masculine

This kind of assignment is interesting because it refers to the difference between natural and grammatical gender. Although this is usually the case, there are examples such as e.g. *muškarčina* “manly man,” or *momče* “little boy” that denote males even though inflect like most feminine and neuter nouns and take feminine and neuter agreement respectively.

**Morphological assignment**

a) Nouns derived with suffix –ina are feminine  
b) Nouns derived with suffix –ost are always feminine.  
c) Nouns derived with suffix –ad are always feminine

### 3.3. Network morphology approach to gender assignment

The goal of this section is to illustrate Network morphology approach to noun inflection classification and gender assignment. It is based on the formal language DATR, which allows computer interpretation.\(^5\) Within this approach *default inheritance* plays a major role.

The basic concept underlining default inheritance is that a word inherits properties from the node that stands in a higher place within the node hierarchy. However, if some property specified in the upper node does not hold for a particular node lower in the hierarchy, the inherited information (default) may be overridden. This can be illustrated with the well-known example related to the bird *penguin*. Namely, penguin is a bird. It means that *bird* stands higher in the hierarchy with regard to *penguin*. One of the characteristics of a bird is that it can fly. A penguin does not have such a possibility, and it would therefore override the default property of the node *bird* (Fraser & Corbett, 1995).

In order to clarify the notion of *default inheritance*, I will introduce the basic concepts of Network morphology.

Namely, default inheritance is defined in Corbett and Fraser (1993, p. 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.”

---

\(^5\) For more on this matter see Brown and Hippisley (2012) or Corbett and Fraser (1993).
The above understanding of default inheritance can be further clarified with the following definitions from Corbett and Fraser (1993, p. 116-118).

Attributes

“An attribute may be atomic or it may consist of a list of atoms. List attributes are descriptions which increase in specificity from left to right.”

Facts

“A fact consists of an attribute: value pair. A value may be stated directly or referenced indirectly by means of another attribute having that value. Chains of reference may be arbitrarily long, though a single attribute may appear only once in any chain. If at the end of a chain of reference, no value can be found for an attribute, the fact in which that attribute appears is undefined.”

Networks

“Lexical information is organized as a network whose basic elements are nodes and facts, and whose structure consists of relationships between basic elements.”

Nodes

“A node is a named location at which one or more facts may be stored.”

Values

“Values may be atomic or list-structured, where a list consists of a sequence of atoms.”

In brief, there are values (contain either atoms or a list of a sequence of atoms, where atoms are considered to be undividable objects), attributes (might be either atoms or may consist of a list of atoms), facts (consist of a pair of attribute and value), nodes (locations where facts are stored), and finally, networks which consist of relationships between nodes and facts (Corbett & Fraser, 1993).
In Figure 1, it is shown how inheritance looks like. Basically, the node *zakon* “law” (with its four facts given: `<stem>`:zakon, `<gloss>`:law, `<nom sg>`:zakon, `<gen sg>`:zakona, with `<stem>`, `<gloss>`, `<nom sg>` and `<gen sg>` being attributes and *zakon*, *law*, *zakon*, *zakona* being their values, respectively) inherits endings from its declension class, which were already defined in the upper node of the declension class in question.

Gender is inferred from either the semantic information of a noun or the declension class it belongs to.

(38) žena:
   <> == NOUN
   <gloss> == woman
   <root all> == žen
   <sem sex> == female


The example in (38) illustrates that semantic information (<`sem sex`> == female) could be sufficient for declension class assignment.

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

The example in (39) illustrates the opposite association. Namely, it shows that gender could be inferred from declension class as well. In this case the noun belongs to the class $i$ (\texttt{<declensional class>} == \texttt{N_{-}I: <mor>}) and therefore belongs to the masculine gender.

An interesting case would be the noun muškarč-in-a which denotes not only “a man,” but “a manly man.” It belongs to the feminine gender, meaning that it takes feminine agreement. However, it denotes a male, and this information has to be specified as well.

(40) muškarč-in-a:
\begin{itemize}
  \item \texttt{<> == NOUN}
  \item \texttt{<gloss> == man}
  \item \texttt{<declensional_class> == N_{-}III : <mor>}
  \item \texttt{<root all> == muškarčin}
  \item \texttt{<sem sex> == male}
\end{itemize}

Finally, the Network morphology approach shows the dynamics between gender and declension class in terms of computational linguistics.

### 3.4. Croatian National Corpus – approach to gender assignment

The traditional grammar, as shown in previous sections, starts describing declension classes from gender. On the contrary, the Network morphology approach acts the opposite way – from a declension class to gender assignment. The approach to gender assignment behind the Croatian National Corpus could be captured as a mixture of these two (traditional and default-based approaches), where traditional grammatical descriptions are well adapted for a computer analysis (Tadić, 1994).

There are two constructs on the basis of which the whole system of noun classification is described:

- sample paradigms (which comprise of common characteristics of a group change)
- a type of change, or a noun class (four different noun classes, which the sample paradigms could be assigned to)

The sample paradigms consist of lemmas (contained in the lexicon), endings and transformations (allowed on stems). Therefore, a list of possible stems, a list of endings and finally the combination rules are established. This makes the system exhaustive, though precise at the same time.

Gender is listed only in case it could not have been predicted from the declension class.
4. Model

This section aims to provide a review of the state-of-the-art in computational modeling of the Croatian noun inflection. This will establish the necessary context of our research. Moreover, the methodology for the gender inducement model (in further text – CroGIM) will be presented.

4.1. Background

The research on computational modeling of Croatian inflectional morphology started before 1990s. The first spelling checker for the Croatian language – Croatian Academic Spelling Checker, emerged in 1994 and it is freely available since 1996. This tool is applied not only e.g. to Microsoft Office, but also to Web search engines (e.g. suggestions posed by Google, such as “did you mean...,” an example taken from Tadić, Brozović-Rončević & Kapetanović, 2012, p. 68). Moreover, based on already existing WordNet for the English language, the Croatian WordNet (CroWN) has been built. This tool deals with retrieving semantically related words. Therefore, with such a tool it is being possible to find web pages that do not contain an exact word but its synonym or otherwise semantically related word (an example given in Tadić et al., 2012, p. 68: nuklearn i atomska energija “nuclear and atomic energy”).

*Croatian Lemmatization Server*, accessible through the database *Croatian MorphologicalLexicon* (cf. Hrvatski morfološki leksikon, n.d.), is another important assistance for the Croatian language. The server was developed at the department of Linguistics Faculty of Humanities and Social Sciences at the University of Zagreb. The number of lexemes the database comprises of (with morphosyntactic tagging applied) exceeds 110,000 (Tadić et al., 2012).

In this thesis an online interface of the Corpus was utilized, particularly the corpus *hnk_v3.0*. The corpus that contains more than 60 million noun forms is automatically morphosyntactically tagged (usually labeled as MSD, where MSD stands for *morphosyntactic description*). The web interface *Bonito2* enabled accessing the corpus v3.0 using queries defined with regular expressions. For instance, an expression `{msd="A.*" & lemma="kuća"}` would result in all the concordances between adjectives and the noun *kuća* “house.”

4.2. Methods

The model given in the Appendices is based on the two different approaches to gender assignment given in section 3.

The model was built in a programming language Python, and the codes used are provided in the Appendices. Rules for the program were obtained by analyzing the descriptive Croatian grammar, reading and combining different formulae, along with the information on the productive morphological patterns found in Radisavljević (2013).

The gender of nouns was meant to be induced on the basis of the following two:

1) Proper noun endings:

Suffixed and unsuffixed nouns have to be differentiated. Why is this important? For instance, both the unsuffixed and the suffixed noun in the following example end in –ost, most “bridge” and radost “joy,” respectively. However, as most unsuffixed nouns that end in –ost, most
“bridge” takes masculine agreement, while radost “joy,” like all the other nouns formed with the suffix –ost, takes feminine agreement:

(41) **lijep most** vs. **lijep-a radost**

“beautiful bridge” “beautiful joy”

The difference between suffixed and unsuffixed nouns is within the scope of this thesis obtained by *listing*. Listing was enabled by the Corpus query system. For example, nouns ending in –ost are mostly derived with the suffix –ost (suffixed), which in turn means that they are mostly feminine. Hence, a small number of the exceptions to this rule (ending in –ost, unsuffixed) are listed. Moreover, there is a noun *kost* “bone” that ends in –ost and it is not derived with –ost, but belongs to the feminine gender. Such an example supports this method of dealing with exceptions, which would otherwise fall under the unsuffixed and thus would be assigned masculine gender.

Listing the exceptions to rules is appropriate especially if a rule is not productive. This is the reason why applying the information on the productiveness of rules is beneficial to this model. The following example of productive –ost suffixation, illustrates a rule: if a noun ends in –ost and it is not in the list, assign feminine gender.

(42) `elif X.value[-1]==‘a’ or X.value[-3:]==‘ost’ and X.value not in [‘most’, ‘post’, ‘gost’, ‘oprost’]:`  

`X.gender.setAll(g.feminine)`

2) Agreement:

Gender of nouns is defined through the agreement with adjectives. Therefore, it is *natural* to utilize the agreement with attributive and predicative adjectives to induce gender. There are two reasons why predicative adjectives are not included. First, there is a practical reason since it was not possible to account for predicative adjectives in the Corpus. Second, there is no extra influence with regard to the agreement, and thus nothing further would be obtained by including predicative adjectives. For inducement of gender through the agreement between nouns and adjectives to be possible, the Corpus was accessed. Therefore, the aim was to build a model that would illustrate the rules people actually use.

Listing the exceptions was obtained by the Corpus query system. Only after listing the exceptions, were the basic rules given to opt for the majority of the examples. This is the way CroGIM will recognize the exceptions.

(43) `[lemma=“.lo” & msd=“N.msn.*”]`

The example of a query in (43) stands for the search for all the nouns ending in –lo that belong to masculine gender, in the nominative singular (“N.msn.*”).

(44) `[lemma = “.*(e|i|u|o)” & msd=“Npfsn.*”]`

The example in (44) is a query of all the lemmas that terminate in one of the given letters (“.*(e|i|u|o)”).
5. Croatian noun inflection with regard to the model

The aim of this section is to show the challenges of the CroGIM with regard to the Croatian nominal inflection and gender system. Therefore, the goal is to understand how the CroGIM grasps the gender system of Croatian and what it fails to attain. First, the focus will be on the difference between proper and common nouns when it comes to the CroGIM. Second, semantic, morphological and phonological gender assignments (as defined in section 3.2) will be presented through the prism of the CroGIM. Afterwards, I will list isolated groups that are not related to any of the discrepancies, but were instead obtained with the corpus queries.

5.1. Proper vs. common nouns

Proper nouns represent a great challenge to the CroGIM. Namely, the CroGIM rules that apply to common nouns do not necessarily apply to the proper nouns. These sections aim at clarifying the disparity between proper and common nouns when it comes to the CroGIM.

Traditional proper male names end in -e, -o, or in a consonant, although male nicknames usually end in –a. Traditional feminine proper names usually end in –o (strong default for feminine nouns). Given that -e, -o and consonants are already marked as productive masculine endings, male names ending in –a impose the only problem. I decided to list those masculine names ending in –a, most of which are nicknames and hypocoristics (in 45 is a piece of the whole code).


X.gender.setAll(g.masculine)

However, the problem remains with the modern proper names recently given to the children. Namely, such names deviate from the traditional ones with regards the CroGIM rules. Particularly, modern feminine nouns often end in –o, -e, -i, or a consonant in addition to the default ending –a. Again, the issue was overcome by listing the exceptions. Still, this is not the best way to deal with proper names, since newly-coined names given after celebrity people, appear every day (e.g. Dolores, Nives, etc.).

Therefore, I decided to list female names that end in –o, -e, -i, or in a consonant. I used queries such as [lemma=",.*(u|x|y|w|q|l|e|o|z|c|a|b|n|m|s|d|f|g|h|j|k|l|p|t|r)" & msd="*Npf.*"] to search the Corpus and find all the feminine proper names (msd=Npf.) ending in any phoneme different from –a (lemma="",.*(u|x|y|w|q|l|e|o|z|c|a|b|n|m|s|d|f|g|h|j|k|l|p|t|r)"").


X.gender.setAll(g.feminine)
Considering the fact that most neuter proper nouns ending in –e are grasped with higher-level rules (on the morphological level), those that do not follow the rules are the only ones to be listed.


X.gender.setAll(g.neuter)

In addition, there are names such as Sani, Vanja, Noa that could refer to both males and females and thus impose a potential problem to the model. This is left to the Corpus to decide.

Furthermore, dealing with surnames, names of cities and villages is in this thesis solved by making a model case sensitive. Hence, the problem with the proper names remains largely open with an aim to be considered properly in the future work.

A further problem represents the automatically tagged Corpus. For example, Nensi “Nancy” is a female name, but it is often tagged as being an adjective – Afpmsny. 6 Shortly, this tag denotes an adjective in its masculine singular form. The context in the Corpus reveals that Nensi “Nancy” most often stands before a surname ending in –i. Since the surname is assigned to the masculine gender, the program presupposes that before that noun stands an adjective that has the same ending.

5.2. Semantic level

According to the semantic assignment rule presented in 3.2, females and males should be assigned feminine and masculine gender respectively. However, the rule does not apply in cases of diminutive forms with the suffix –če, or augmentative forms with the suffix –ina.

The examples in (47) and (48) illustrate CroGIM’s dealing with the cases of diminutives. The sex such nouns denote is added as extra information.

(48) elif X.value in ['dečače', 'momče']:

X.gender.setAll(g.neuter)
X.gender.semantic(g.masculine)

(49) elif X.value in ['devojče']:

X.gender.setAll(g.neuter)
X.gender.semantic(g.feminine)

The augmentative suffix –ina is a productive feminine suffix (Radisavljević, 2013). Examples of augmentatives denoting males are listed.

---

(50) `\text{elif X.value[-3:] == 'ina' and X.value in ['budaletina', 'detektivčina', 'dimničarina', 'djeđetina', 'djeđurina', 'dječačina', 'gazdetina', 'generalina', 'glaveština', 'ljudina', 'muškarčina']:}
\begin{align*}
\text{X.gender.setAll(g.feminine)} \\
\text{X.gender.semantic = g.masculine}
\end{align*}

\text{elif X.value[-3:] == 'ina':}
\begin{align*}
\text{X.gender.setAll(g.feminine)} \\
\end{align*}

Moreover, as already mentioned in section 2.1, nouns such as \textit{pijanica} “drunkard,” \textit{varalica} “cheater” could refer to both males and females, and usually take feminine agreement in both cases.

In addition to these two suffixes, suffix \textit{–lo} is a productive neuter suffix (cf. Radisavljević, 2013) (e.g. n. \textit{pisa-lo} “an item used for writing,” n. \textit{beli-lo} “bleach”). Such neuter nouns are built from a verb denoting an action of e.g. writing or bleaching and the suffix \textit{–lo}. Moreover, it is worth mentioning that in the same way as used for building nouns denoting objects, verb plus \textit{–lo} could be applied to forming nouns referring to humans, both females and males. In any case, taking neuter agreement is common. However, many of those examples found in Stevanović (1970) are not found in the Corpus. Even if they are found in the Corpus, such examples are often wrongly tagged. For example, n. \textit{zakera-lo} “nagger” is tagged as a verb. Therefore, for the purpose of this thesis, the issue regarding nouns ending in \textit{–lo} will be handled in the following way: all the exceptions to the general rule stating that nouns ending in \textit{–lo} are neuter, are listed (example in 51).

In addition, as Stevanović (1970) noted, neuter nouns ending in \textit{–lo} have various other meanings (e.g. there are old-established nouns ending in \textit{–lo}, e.g. n. \textit{selo} “village” or n. \textit{telo} “body,” or male names formed by adding the suffix \textit{–lo}, such as m. \textit{Momči-lo} from n. \textit{momč-e} “boy,” m. \textit{Dobri-lo} from f. \textit{Dobra}).

(51) `\text{elif X.value[-2:]=='lo' and not X.value.istitle() and not X.value in ['a-raphael-tiepolo', 'diavolo', 'discelpolo', 'escenabolo', 'karlo', 'mullo', 'mulo', 'polo', 'rolo', 'vaterpolo'] or X.value in ['Etnoselo']}:}
\begin{align*}
\text{X.gender.setAll(g.neuter)} \\
\text{elif X.value[-2:]=='lo' and X.value.istitle():} \\
\text{X.gender.setAll(g.masculine)}
\end{align*}

\textbf{5.3. Morphological level}

The aim of this section is to describe how CroGIM deals with productive and default suffixes in Croatian. In particular, how it is dealing with the productive or default suffixes that violate phonological defaults or productive endings. For example, masculine suffixes that terminate in –
and feminine suffixes that terminate in a consonant are in concordance neither with default endings for the masculine and the feminine gender (consonants are default masculine endings, and default feminine ending in –a) nor with the productive masculine ending –o/-e.

The suffix –ša (along with suffixes –eša, -iša, -oša) is commonly used for building masculine hypocoristic forms and masculine names. Therefore, exceptions to this rule are listed, and afterwards the general rule is outlined.

(52) elif X.value in ['kiša', 'piša', 'poliša', 'gejša', 'niša', 'čaša', 'kaša', 'snaša', 'šalša'] or X.value in ['Urša', 'Maša', 'Deša', 'Ajša']:
    X.gender.setAll(g.feminine)
elif X.value [-2:] == ‘ša’:
    X.gender.setAll(g.masculine)

The similar is done with the suffix –elja. The suffix is not productive, but it is the default masculine suffix. Therefore, nouns not derived by the suffix –elja belong to feminine gender and are excluded from the general rule stating that nouns ending in –elja are masculine.

(53) elif X.value in ['želja', 'povelja', 'nedelja', 'znatiželja', 'kecelja', 'nedjelja', 'felija'] or X.value in ['Zdelja', 'Sv.Nedjelja', 'Ljelja', 'Lelja']:
    X.gender.setAll(g.feminine)
elif X.value [-4:]=='elja'
    X.gender.setAll(g.masculine)

The suffix –onja is a default masculine suffix. Nouns derived by this suffix usually have negative connotation, e.g. rog-onja “cuckold.” Therefore, only the exceptions are listed.

(54) elif X.value in ['Sonja', 'Dragonja', 'Ronja', 'Tonja', 'Monja', 'Dragonja', 'Burgonja', 'Lonja', 'Bizonja', 'Hercigonja', 'Letonja', 'Monkodonja', 'Osonja', 'Bolonja', 'Slivonja', 'Zelina-Lonja', 'Mirna-Dragonja'] or in ['bolonja']:
    X.gender.setAll(g.feminine)
elif X.value [-4:]=='onja'
    X.gender.setAll(g.masculine)

The suffix –če is a productive suffix used in deriving diminutive and hypocoristic forms (Radisavljević, 2013). It is especially common for building nouns denoting names of offspring’s or young human beings (Stevanović, 1970).

(55) elif X.value [-2:]=='če' and not X.value.istitle():
    X.gender.setAll(g.neuter)
The suffix –ce, together with suffixes -ance, -ence, -ašce, -ešce (of which all end in –ce, and thus allow for grasping in one single rule), is a suffix used for deriving diminutive and hypocoristic forms. The only exception to this rule is a group of non-traditional proper male names. Interestingly enough, only the suffix –ce is a productive one, all the others (-ance, -ence, -ašce, -ešce) are not.

(56) elif X.value in ['Bece', 'Blace', 'Broce', 'Bruce', 'Bryce', 'Calace', 'Clarence', 'Enface', 'Felice', ...]:

    X.gender.setAll(g.masculine)

    elif X.value [-2:] == 'ce':

    X.gender.setAll(g.neuter)

The suffix –ost is productive for derivation of feminine nouns, with only few exceptions provided in a list. Although already outlined earlier in section 4, it is again exemplified in (57).

(57) elif X.value[-3:]=='ost' and X.value not in ['most', 'post','gost','oprost']:

    X.gender.setAll(g.feminine)

The suffix –ad is a productive feminine suffix (cf. Radisavljević, 2013). Those unsuffixed nouns that end in –ad are listed as exceptions for the generalization – nouns ending in –ad are feminine.

(58) elif not X.value.istitle() and X.value[-2:]== 'ad' and not X.value in ['antikomad', 'autootpad', 'bead', 'bio-otpad', 'brodovad', 'cestamad', 'džihad', 'džihad', 'džihad', 'fad', 'gad', 'grad', 'had', ...]:

    X.gender.setAll(g.feminine)

Suffixes –lija and –džija are masculine suffixes adopted from the Turkish language. It is an interesting fact that these suffixes were not productive for a period of time. Nowadays, the suffix –džija is again productive for deriving nouns denoting executors of different actions. However, it often indicates a marked connotation (Radić, 1999).

(59) elif X.value in ['Neimarlija', 'Nišlija', 'Sarajlija', 'Bešlija', 'Laklija', 'Šarlija', 'Vujaklija', 'Budimlija',...]:

    X.gender.setAll(g.masculine)

    elif len(X.value)>4 and X.value [-5:] in ['alija', 'elija', 'ilija', 'olija', 'ulija']:

    X.gender.setAll(g.feminine)

    elif X.value [-4:] in ['džija', 'lija'] and not X.value.istitle():

    X.gender.setAll(g.masculine)

Most neuter nouns ending in -e could be embraced by the suffix –je, along with its different allomorphs (-anje, -ljje, - će, -de).
X.gender.setAll(g.masc)

elif X.value [-3:] in ['če', 'nje', 'lje', 'žde']:
X.gender.setAll(g.neuter)

elif X.value [-2:] == 'je':
X.gender.setAll(g.neuter)

The suffix –ište is a productive neuter suffix for the construction of nouns referring to a place.

X.gender.setAll(g.neuter)

elif X.value in ['misao', 'pomisao', 'izrastao', 'so', 'mati']:
X.gender.setAll(g.femin)

5.4. Phonological level

Neuter nouns in –e/-o/-u/-i that did not fall under the morphological level and are violating the general rule that nouns ending in –e/-o/-u/-i are masculine, were listed due to the fact that those are not productive (Radisavljević, 2013).

X.gender.setAll(g.neuter)

Masculine nouns ending in a consonant are productive (Radisavljević, 2013). Since it would be neither possible nor efficient to list all those nouns, I put the general rule in the model, where nouns ending in a consonant are masculine. The rule applies only in the cases when the search passes by the list of feminine nouns terminating in a consonant.

elif X.value in ['art-krv', 'bojazan', 'bit', 'bol', 'bolest', 'dobit', 'draž', 'gemišt-vlastčar', 'glad', 'golijen', 'kap', 'kob', 'korist', 'krv', 'kčer', 'laž', ...]:
def X.gender.setAll(g.feminine)

eelif X.value[-1] in ['b', 'v', 'g', 'd', 'z', 'ţ', 'j', 'k', 'l', 'p', 'r', 's', 't', 'ć', 'f', 'h', 'č', 'š']:
    X.gender.setAll(g.masculine)

5.5. Isolated groups

There are nouns such as *kino* “cinema” that change their gender in the plural. Those nouns are listed and their gender is marked separately for the singular and the plural.

(65) elif X.value in ['kino', 'dinamo', 'torzo', 'libreto', 'molo', 'veto', 'salto', 'korzo', 'geto', 'torero', 'mikado']:

    X.gender.singular = g.masculine
    X.gender.plural = g.neuter

A noun *doba* “era/period” is the only neuter word in –a.

(66) elif X.value in ['doba']:

    X.gender.setAll(g.neuter)

5.6. Regular groups

There are regularities acquired by means of the Corpus query system.

Namely, by analyzing different queries, it could be easily inferred that nouns ending in -ko, -ao or -go are mostly masculine. This is the reason why I decided to list the neuter ones (example 67).


    X.gender.setAll(g.masculine)

The number of masculines found in –vo is rather small. Therefore, before the generalization “all nouns ending in –vo are neuter,” the exceptions are listed (example 67).


    X.gender.setAll(g.masculine)

eelif X.value[-2:]=='vo':
X.gender.setAll(g.neuter)

5.7. Agreement with adjectives

The second part of the model refers to binding the code in Python with an online interface of the Corpus in order to get the agreement of a given noun with adjectives. The whole Python code was translated into the programming language JavaScript in order to be more easily implemented (see Appendices). Knowing that it is adjectives that define gender of nouns, connecting the model to the Corpus seems to be the best solution for gender assignment.

The idea implemented in the model was that adjectives in the singular form ending in –a in are feminine, adjectives in the singular form ending in –e or –o are neuter, and those ending in a consonant or in – i are masculine. The query I decided to use for eliciting the adjectives that stand before the noun of which we want to find out the gender is [msd="A.p.sn.*"] [lemma="auto" & msd= "N..sn.*"]. Here, auto “car” stands for any noun the user is searching for. Being aware of possible tagging errors in the Corpus, I decided to conduct the statistics based on few examples in order to make sure that the right gender would be attributed. For instance, the output of the query for the noun auto “car” is shown in Figure 2.

![Figure 2. The agreement of the noun auto “car” and its adjectives](image)

In the third example, the adjective, as well as the noun auto “car” is in the genitive singular form, while their tags suggest that both the adjective and the noun are feminine and singular. Certainly, ending –a, strong default feminine ending, is responsible since it drew the program to assign feminine gender.

In addition, there are examples of adjectives that agree with masculine nouns and end in –o, which are the examples of l-vocalization in Croatian. L-vocalization refers to the process by which historical –l in final position changed to –o. These adjectives are formed by the process of adjectivization (conversion of participles into adjectives, as mentioned in section 2.2.2).

E.g. debeo dječak “fat boy,” where the old from debel changed to debeo. The facilitating factor regarding the CroGIM is that such adjectives have a vowel before the final vowel –o. Therefore, putting it as a rule, it narrows down the possible errors.

---

7 There are few examples of adjectives ending in –e, e.g. sv-e vrem-e “all the time.”

6. Cognitive aspects of noun inflection in Croatian

The aim of this section is twofold. First, to indicate the complexity of Croatian nominal morphology, the complexity that makes it interesting for cognitive science. The second aim is to give an overview of how gender is acquired by children (section 6.1), given that language acquisition could be viewed as a field at which different theories in cognitive science could be tested (e.g. Pinker, 1995).

Merging semantic and formal rules makes inflection of nouns attractive to cognitive science. In Croatian, for example, the noun *dijete* “child” is neuter in the singular and belongs to the class a (see Table 1), while in the plural (*djeca*, “children”) it inflects like most feminine singular nouns and belongs to the class e. Additionally, the noun *drvo* “tree” also shows different morphological behavior with regards to whether it denotes a living tree or a construction material. With reference to the Corpus, it is interesting to note that the longer form in genitive *drveta* i *obojen temperom*, “…it is made of tree and painted with tempera.” Moreover, there are indications that having formal gender affects cognitive processing. In particular, children acquiring languages having gender grammatically expressed might be aware of gender much earlier than those children not having it in theirs (Flaherty, 2011). Namely, “Guiora et al. (1982) suggest that children learning Hebrew, which has grammatical gender, come to recognize their own gender identity earlier than those learning English (in which gender has a minor role) or Finnish (which has no gender category)” (Flaherty, 2011, p. 30).

Furthermore, with regards to the computer analysis, it could be inferred from the previous sections how much effort is required for the computational description of one single piece of nominal inflection – gender. Inflection is intuitive for each native speaker. Therefore, in terms of computer-mind relationships, we come to the well-known question of whether human intelligence could be modeled.

6.1. Language acquisition

*The richer noun or verb morphology is in the input, the more stimulated the child will be to develop noun or verb morphology rapidly.*

(Dressler, 2007, p. 8)

The field of language acquisition is of great importance for cognitive science and it is no surprise that theories engaged in understanding human cognition often struggle to explain language acquisition (cf. Baudoin de Courtenay, 1974 and Jakobson, 1941, 1977 as cited in Dressler, 2012). Why is it so? As Bowerman and Levinson (2001) stated, cognitive and linguistic development research flourished separately, as the relationship between the two was not found in the beginning. First, their methods were different. Namely, research on cognitive development is still carried out mainly by conducting experiments on children of different ages. On the contrary, research on language acquisition is often done by observing children in a natural environment. Moreover, the period of Chomskyan revolution from the year 1957, when Chomsky published his book “Syntactic structures” had an influence on conceiving a language as a unique part of the cognitive system. In addition, the stages of cognitive development Piaget (1973) established were not in congruence with the language acquisition data.
How do cognitive development and language acquisition relate? MacWhinney (2010) in his paper classifies all current theories into two main paradigms. Generative grammar constitutes one of them. The main generativists’ idea is the existence of one innate grammar behind all existing languages, which is the concept of Universal Grammar (UG). The other paradigm consists of all the theories that assume language is developed by virtue of natural processes in the brain, body and/or environment. Therefore, this paradigm combines the knowledge obtained from cognitive linguistics, functional linguistics, embodied cognition, neural network modeling, corpus analysis, cognitive neuroscience in order to investigate biological and statistical means that create language structure (as cited in MacWhinney, 2010, p. 20). Some of the most important differences between the two paradigms arise from the different view on the relationship between language and cognitive system in general. According to the first one, language evolved as a unique ability of humans that cannot be acquired only from the input, instead there must be certain genes that give rise to language development. On the other side, the second paradigm views language as any other cognitive ability. This in turn means that language acquisition is possible without any special genes.

For instance, Neo-Chomskyan view cognitive development as “necessary but not sufficient condition for semantic development” (Gopnik, 2001, p. 50). On the contrary, the work of Piaget (as cited in Gopnik, 2001), who made a great contribution to research on cognitive development, falls into the second paradigm. Piaget (as cited in Gopnik, 2001) thought cognitive development precedes language development. Moreover, there is the classical interactionist view, e.g. by Whorf or Vygotsky, both of whom suggest that language plays a crucial role in shaping cognition (cf. e.g. Gopnik, 2001). A question that naturally arises is how language and cognition are interrelated during the acquisition period.

Psycholinguists are putting much effort into developing a theory that would explain the relationship between language and cognition. They are changing methods and approaches in order to obtain better insight into the bond between them. In particular, this means changing the focus from structural to the semantic content children acquire and finding its relationship with non-linguistic categories (cf. Bowerman & Levinson, 2001). By focusing on the concepts of e.g. space, time or numbers, researchers try to find out whether children understand such concepts before they even start to use words related to them, i.e. whether the understanding of these non-linguistic categories develops separately from the words that relate to them. Moreover, whether the usage of words related to these concepts flourishes independently of a specific language, or they construct it through the language and culture they belong to (Bowerman & Levinson, 2001). In addition, for the purpose of clarity it has to be made a distinction between the child-directed speech, adult-directed adult speech and a target language.

Now that the relationship between language acquisition and cognition is less obscure, on one side I want to present the acquisition of inflectional morphology with regard to several different theoretical approaches. On the other side, I wish to review the knowledge of the acquisition of gender in Croatian. These two aims will be the topics of the following subsections.

6.1.1. Acquisition of inflectional morphology

Inflectional morphology has an important place in language acquisition theories. Moreover, from mid 80s inflectional morphology has been at the center of the debate between the two important approaches to human cognition – the symbolic and the anti-symbolic approach (Penke, 2012).
While the symbolic approach to processing supports the claim that inflected words are “composed out of component morphemes by application of a mental operation that combines morphemes displaying the right abstract features such as [+V] or [+PAST]...anti-symbolic approach assumes instead that inflected forms are structurally non-compositional and are learned and stored as whole-word forms in an associative memory network” (Penke, 2012, p. 2).

This initial debate was followed by papers supporting one or the other view of cognitive processing. Proponents of the dual-mechanism model, one of the symbolic approaches, supported e.g by Chomsky and Pinker (cf. e.g. Argus, 2012), assume that irregular forms are unanalyzed and memorized as a whole. However, within this route the analogy plays an important role. Namely, irregular forms are thought to be stored in clusters, where phonological similarity plays an important role in the association of the instances. For example, patterns like ring/rang/rung, sing/sang/sung, spring/sprang/sprung would make speakers guess the past tense for similar verbs by analogy. The term coined by Motsch (1981 as stated in Dressler & Laaha, 2012) for this type of analogy where “a neologism [is] formed in exact imitation of one specific existing term” is surface analogy. The term analogy dates back from the mathematics in the age of Ancient Greeks, where it denoted a mathematical proportion. Afterwards, the term was adopted by the grammarians Aristophanes of Byzantion and Aristarchos of Samothrake who wanted to apply mathematical proportions on morphological patterns. Interestingly, Roman grammarians added some other meanings to this term: “(similium) comparatio, proportio, secundum rationem, analogia, but also stronger regula and weaker similitudo” (Dressler & Laaha, 2012, p. 48). This explains why nowadays we use analogy in many different disciplines and contexts.9 However, as Dressler & Laaha (2012, p. 48) noted: “Just one other term of Greek origin is missing: schema.” As Dressler and Laaha (2012) noted, surface analogy cannot be detached from morphology. Viennese acquisition model (cf. Laaha et al. 2006, Korecky-Kröll et al. 2012 in Dressler & Laaha, 2012) makes a distinction between productive rules, unproductive rules and schemas (“relevant for isolated paradigms and families of paradigms” as stated in Dressler and Laaha, 2012, p. 49). Here, rules are not understood as strict commands, but as patterns that differ with regard to productivity level. Their model shows that surface analogies are not related to isolated patterns and that most often analogies are related to productive patterns. As for the unproductive ones, they did find the usage of irregular brang instead of brought in the child speech, formed by analogy to the subregular rule or schema (sing, sang, sung). However, as they noticed, there were no inverse analogical changes e.g. wrang to *wrought (Dressler & Laaha, 2012, p. 56). The conclusion they arrived at is that the productivity of a rule affects the number of analogies that would occur during the acquisition. Moreover, it cannot be that surface analogies happen in the analogy to stored inflectional forms as such, meaning that some level of abstraction to either an unproductive rule or a schema is necessary (Dressler & Laaha, 2012). The only exceptions to this, according to Dressler and Laaha (2012), are extragrammatical surface analogies, related to the period of language acquisition when grammatical modules have not yet been developed (see section 6.1.1).

---

9 For example, analogy is often used in cognitive science in general (see e.g. Dressler & Laaha, 2012; Chalmers, French & Hofstadter, 1991).
Constructivists describe language acquisition through different developmental stages but also recognize that language structures do not act independently. Argus (2012) considers the constructivist approach different from and superior to the three approaches mentioned since constructivists assume that “purely morphological and language-specific factors play a secondary role in the acquisition of inflection” (Argus, 2007, p. 16). Namely, according to the constructivists acquisition of inflectional morphology starts later than lexical, phonological and syntactic acquisition, but earlier than the acquisition of derivational morphology (except for diminutives and/or hypocoristics) (Dressler, 2003). Natural Morphology is compatible with constructivism. It seeks for the natural patterns in morphology and language acquisition is along with language processing and the study of language impairments reasonably highly important field within the scope of Natural Morphology. These are the fields where the core concepts such as productive, natural and unmarked could be deepened (Radisavljević, 2013). Therefore, a pattern starting from universal preferences (first subtheory), through typological, to language specific morphological processes is a pattern also found in acquiring a language. Namely, children in their early phase utter many extragrammatical formations (e.g. reduplications, back-formations, echo-words). Afterwards, they become aware of typological and language-specific properties (Dressler, 2005; Radisavljević, 2013).

According to the view of Natural Morphology, there are three different stages noted in the development of morphology: pre-morphology, proto-morphology and proper morphology period. In the first, pre-morphology period, there is no morphology. It is the period when the acquisition of morphology is related to the general cognitive principles since grammatical modules have not developed yet. Within this period, the production of extragrammatical formations (such as e.g. reduplications, back-formations, echo-words) is usual. Furthermore, in this phase children memorize words without analyzing. In the next phase, protomorphology phase, children begin to analyze words. Hence, it is the period when most overgeneralizations occur. Overgeneralization is one of the consequences of the fact that children identify rules and conceive them as being productive, be they productive or not (cf. Dressler, 2003). Therefore, the utterances are being analyzed, and many wrongly formed words are produced due to the analogy with productive defaults. The evidence suggests that the morphological analysis starts at the two-word stage (cf. e.g. Dressler, 2003; Penke, 2012). The beginning of this phase is defined with the occurrence of the first mini-paradigms – referring to at least three different inflectional forms of the same lexeme. In this phase first analogies and association rules are made. It is assumed that this period ends when grammatical modules are developed, which is at the same the beginning of the period of proper morphology.
6.1.2. The acquisition of gender in Croatian

The aim of this section is to review the findings on gender acquisition in the Croatian language. As already mentioned above, there is no much difference between Serbian and Croatian when it comes to nominal inflection (cf. Kovačević, Palmović & Hržica, 2009). Therefore, I find the studies done on Serbian language quite relevant and valuable for the thesis. There are two relevant and interesting studies that refer to the gender of nouns in Croatian. The first refers to acquiring nominal morphology in general (Kovačević et al., 2009) while the other one refers to acquiring diminutives in Croatian (Palmović, 2007). Moreover, there are two studies that relate to the role of diminutives in acquiring gender in Serbian, namely Ševa (2006) and Ševa et al. (2007).

Kovačević et al. (2009) researched the acquisition of nominal morphology in Croatian. Gender was explored in addition to case and number. The focus of the study was on the implications of morphological richness to acquiring inflectional morphology. In particular, the interval for acquiring particular noun inflection categories – case, number, gender. The study was based on the data obtained mainly recording the speech of the girl Antonija between the ages 1;3 and 2;8. However, the data from the child language corpus were taken into account for the comparisons (Kovačević et al., 2009).

With regards to gender, Antonija used feminine nouns in her speech more often than masculine and neuter (58% feminine, 36% masculine, neuter 6%). This could be explained to a certain extent in terms of contexts of conversations while it could be also explained by the transparency of the feminine paradigm. Neuter nouns were restricted to only a few contexts, usually related to food and drink, limited to only a few lemmas. Nouns were frequently diminutivized. Unsurprisingly, the proportion of feminine, masculine and neuter nouns Antonija used reflects the child-directed speech.

In addition, applying diminutives can make the paradigms more transparent. For example, as Palmović (2007) reported, number of diminutives have more transparent paradigms compared to their simplex forms – e.g. ruk-ic-a “hand-DIM” vs. ruk-a “hand,” nog-ic-a “leg-DIM” vs. nog-a “leg,” knjig-ic-a “book-DIM” vs. knjig-a “book;” dative/locative forms: ruk-ic-i vs. ruč-i, nog-ic-i vs. noć-i and knjig-ic-i vs. knjig-i (p. 82). Their simplex forms are less transparent in the oblique cases due to the shift called sibilisation, where k, g, h change to c, z, s, respectively. Moreover, diminutives reduce the number of noun classes. Nouns that belong to class i (feminine nouns ending in a consonant), which is a weakly productive class, derive diminutives with –ica, likewise to nouns that belong to class e. Therefore, both nouns belonging to the class i and nouns belonging to the class e derive diminutives that belong to the class e (most feminine nouns ending in -a). What is important is that Antonia preferred diminutive forms, even though they usually have longer form when compared to simplex nouns, and they require some derivational rules. Moreover, after the age 2;0, she used simplex forms in nominative, but in oblique cases she used diminutive forms. For instance, leptir “butterfly,” leptirića “butterfly-DIM-GEN” (Palmović, 2007, p. 81).

The range of diminutive meanings is wide, from denoting smallness to pejorative meaning and irony, depending on the word that is diminutivized or on the context. However, Palmović (2007) concludes that Antonija did not use diminutives to express smallness and agreed with Dressler and Karpf (1995) on that their early emergence is subjected to the pragmatic reasons. Palmović
(2007) wrote that the function of diminutives in the early period is “to create a child-oriented situation or the specific atmosphere of endearment and love” (p. 85).

Since studies that involved Russian and Lithuanian revealed that diminutives are quite common in child language, Ševa (2006) examined the role of diminutives in Serbian language with regard to acquiring noun morphology. She concluded that in Serbian, as well as in Russian and Lithuanian, diminutives happen to be of importance for facilitating gender acquisition. Often, it is the first morphological rule children acquire (cf. Ševa, 2006; Savickiené & Dressler, 2007). Moreover, as stated in Ševa et al. (2007), it was found in previous studies that Polish and Russian children make less case-marking errors with diminutives when compared to simplex nouns. Interestingly, both in Polish and in Russian diminutives are in child-directed speech used much more frequently than in Serbian (Ševa et al., 2007). This was the reason why Ševa et al. (2007) were interested to explore whether the advantage of diminutives was only due to the high frequency of use or diminutives do have an important role in acquiring nominal morphology. The former would be diminished in case diminutives prove to be facilitators for gender acquisition in Serbian as well.

In Serbian, as well as in most Slavic languages, diminutives preserve the gender of the simplex noun, and the suffixes for diminutives are “transparently gender marked” since from their ending it is easy to infer about the gender (Ševa et al., 2007, p. 114):

(69) 

\[ \text{zanimljiv-a knjig-a} \]

- interesting-Fem.N.Sg book-Fem.N.Sg
- “interesting book”

\[ \text{zanimljiv-a knjiž-ic-a} \]

- interesting-Fem.N.Sg book-Fem.N.Sg.Dim
- “interesting bookDim”

Ševa et al. (2007) concluded that in Serbian, as in Russian, children were better in performing gender agreement with diminutives when compared to simplex nouns.

The only difference between Russian and Serbian children appeared to be the difficulty of rejecting feminine pseudo nouns (ending in -a) for Serbian children. Namely, as part of the experiment, the children were introduced to unfamiliar objects and animals, half of which were named with pseudo-forms. It was concluded that both Russian and Serbian children perform better with familiar compared to unfamiliar nouns. Moreover, they also perform better with diminutives than with simplex word forms. Therefore, Ševa et al. (2007) gives two possible explanations for why feminine pseudo nouns were difficult for Serbian children. One such explanation would be that in Serbian there are numerous hypocoristic forms for masculine animals and kinship terms ending in -a which may additionally confuse children in gender agreement task (e.g. medved “bear” vs. meda “teddy bear,” and although such examples exist in Russian as well, e.g. misa “teddy bear” or zajka “little rabbit,” in Serbian these hypocoristic forms are more productive and more frequent, as stated in Ševa et al., 2007, p. 125). The other possible reason for the difference between Serbian children and Russian children with regards to the gender effect with feminine novel nouns, is also a possibility that such a slight difference is a mere consequence of the selection of the nouns to be novel for Serbian, given that these are adjusted Russian novel nouns. Moreover, it is confirmed as well (when these nouns were
compared individually) that especially timza and mompa (novel nouns) elicited more errors in Serbian than in Russian, which suggests that these might be simply less acceptable pseudo-words in Serbian (Ševa et al., 2007, p. 125).

Finally, this difference between Serbian and Russian children does not diminish the great advantage of diminutives for gender-agreement production, even though the frequency of diminutives in Serbian child-directed language is considerably lower (cf. Ševa, 2006). Ševa (2006) also notes that Serbian parents have a tendency to use nearly ten times more diminutives in conversation with their children compared to the conversation with adults. In spite of the fact that it is still remarkably fewer (which is according to Ševa (2006) 7% of diminutives out of the total number of noun tokens and 11% out total number of lemmata) if compared to the total number of diminutives Russian, Polish or Lithuanian parents use (according to Ševa (2006) 20-45% of diminutives). It is interesting to note that diminutive productivity is similar in those languages (cf. Ševa, 2006).

Furthermore, there is a study that approached noun gender from a different angle. Namely, the study by Radanović and Milin (2011) aimed to explore whether the presence of semantic information about gender is relevant as well in the tasks not directly related to gender processing, such as e.g. lexical decision task.

The main hypothesis was that due to the transparency of the natural gender for animate nouns with sibling in the other (e.g. lav “lion” – lavica “lioness”), there was a match between semantic and grammatical information. This match was believed to facilitate the processing of the nouns. Vice versa, not having a sibling in other gender should make the processing harder (Radanović & Milin, 2011). Their experiments confirmed the assumptions, and showed that animate nouns with a sibling in the other gender were processed faster than those without a sibling (e.g. mornar “sailor” or žirafa “giraffe”) (the examples are taken from Radanović & Milin, 2011, p. 353). Moreover, nouns with a sibling were processed faster than inanimates as well, with only the subtle difference in processing. They presume that natural gender could facilitate word processing, since neither animates without their sibling nor inanimates are marked for natural gender. It has already been demonstrated in earlier experiments that in tasks which directly involve grammatical gender processing, the information about natural gender indeed facilitates processing (e.g. Schiller et al., 2003 as cited in Radanović & Milin, 2011). However, this study showed it is likely that this information is also used in a task which does not require grammatical gender processing, but only the meaning of a word.

The question that is left open in Radanović and Milin (2011) is concerned with inanimates semantically marked for gender. It is probable that gender is included in the representation of such nouns, and therefore, these nouns might be more quickly processed compared to non-semantically marked inanimate nouns. Hence, it is yet to be demonstrated if having a sibling in the other gender is a facilitating effect for processing animate nouns, or such an effect is recognizable in inanimates as well.

In addition to all these studies, there is a recent account on adjective and semantics acquisition in Slovenian (Petrič, Ljubič, Oblak, Korecky-Kröll & Dressler, 2014). Namely, they studied one Slovenian girl and it was demonstrated that she rarely used neuter gender. Neuter gender appeared only at the age of 2; 3 (Petrič et al., 2014). Interestingly, as they pointed out, even the noun otrok “child” in Slovenian is masculine. The girl mostly used feminine gender, more than
twice more than masculine. This was attributed partly to the father’s speech directed to the child and also to the topic of their speech – which were colors. Even the noun *barva* “color” is feminine as well as the nominalized forms of color that are often used (Petrič et al., 2014).

To conclude, gender of nouns could be an important facilitator for the acquisition of the Croatian nouns (Ševa et al., 2007) as well as for the noun processing (Radanović & Milin, 2011).

6.2. Gender – arbitrary or semantically motivated?

One cannot think about gender without posing a question such as why e.g. *auto* “car” is masculine in Croatian, and neuter in Deutsch. Is it possible that grammatical gender is an arbitrary category? Furthermore, to which extent is gender arbitrary or not? The question of grammatical gender and its arbitrariness dates back to the time of Ancient Greece. Sextus Empiricus (according to Blank, 1998 as cited in Kilarski, 2007) “considered presence of epicens”\(^{10}\) and assigning f. or m. gender to inanimates as examples of the anomalies in the relation between natural and grammatical gender” (p. 2).

It might be that the reason why we have difficulty understanding the notion of gender is in terminology. Namely, we are often confronted with terms such as *natural* and *grammatical* gender. This makes us assume there are two kinds of gender. However, these could be perceived as two perspectives of gender. As Nikolić (2002-2003) put it, we can see gender as a lexical-semantic category in which case there are nouns that have natural gender and those that do not have it. Moreover, we can view gender as a grammatical category in which case there are three genders: feminine, masculine and neuter (determined through the agreement with adjectives).

It is a common view that so-called natural gender is semantically motivated. This is the most expected with humans, who are on top of the animacy scale, less so for animals, much less in plants and none in inanimates. Namely, most nouns denoting males are masculine, while those denoting females are feminine. Although, it is not always true. For example, German *Mädchen* “girl” and Croatian *devojče* “girl” both belong to the neuter gender, although they refer to females. Why is this so? According to Toyota and Vlasa Florea (2009), in order to understand such examples we have to understand the diachronical perspective of gender. Namely, Proto-Indo-European (PIE), about 6000 years ago, had a binary gender system distinguishing between *active* and *inactive* nouns. Active nouns had the ability to initiate an action and thus animates belonged to this group, while inactive nouns did not have such a possibility and thus consisted of inanimates. Later, in Indo-European languages that evolved from PIE, feminine and masculine gender developed from the active gender, while neuter gender evolved from the inactive gender.\(^{11}\) The explanation why e.g. *Mädchen* is neuter, as stated in Toyota & Vlasa Florea (2009) would be that it refers to a young feminine person still not being able to reproduce, since she still has not reached sexual maturity. Consider the cases in concordance with this from Toyota and Vlasa Florea (2009):

---

\(^{10}\) Epicens are nouns that can refer to both males and females (e.g. *pijanica* “drunkard”).

\(^{11}\) There were also exceptions to this gender system based on animacy, see more in Toyota and Vlasa Florea (2009).
In southern dialects of Polish, the neuter is employed when the referent is an unmarried girl. After the marriage the gender changes to the feminine one. In Czech and Slovak, the female person is also referred to as neuter unless she is married (Corbett 1999: 100). In Konkani (Indic), words such as bayl “woman” andawoy “mother” have double gender: they have the agreement with the feminine gender when used to older women and the neuter when referring to young women. According to this system, Konkani has two separate feminine pronouns, e.g. tē refers to a young woman and ti, to an older woman. (p. 169)

In addition to this, most problematic animates are nouns that can refer to both males and females (e.g. pijanica “drunkard”), nouns that could have two genders (e.g. bol “pain”), nouns that change their gender in plural (m. sluga “servant” in plural f. slug-e, n. oko “eye,” in plural f. oč-i, the same principle is for n. uho “ear,” pl. f. uš-i), nouns that could have metaphorical meaning (e.g. kukavica “cuckoo,” also can refer to a coward, a person who does not have enough courage) (see Tafra, 2001, p. 254).

Nevertheless, inanimates represent a true problem when it comes to the semantics of gender. It is much harder to trace the nature of grammatical gender of inanimates. Therefore, it is often attributed to arbitrariness. Arnauld (1612-1694) and Lancelot (1615-1695) considered gender of animates to be semantically motivated, while gender of inanimates was considered arbitrary. However, “according to a succession of scholars (Herder 1772, Adelung 1783, Humboldt (e.g., Humboldt 1827), and Grimm 1890), the origin of the category was sought in imagination and personification, in an extension of natural gender to inanimates, based on semantic oppositions such as active vs. passive” (Kilarski, 2007, p. 3). On the contrary, although Brugmann (1889 as cited in Kilarski, 2007) did not negate the connection between natural and grammatical gender, he did negate the likelihood that the gender of inanimates is semantically motivated by the mechanism of personification.

Interestingly, Baudouin de Courtenay (as cited in Kilarski, 2007) not only thought grammatical gender to be a sort of mistake or an accident, but he also compared it to a “deformity or a bad habit, and held it responsible for a range of human afflictions, including nightmares, pathological behavior, erotic and religious delusions and sadism” (p. 5). On the contrary, there was as well the stream of scholars (e.g. Grimm, 1890 as cited in Kilarski, 2007) considering gender to be a sign of language development.

Back to the present time, in many languages there could be made generalizations such as e.g. in German, wh car names are usually masculine, while ship names are feminine (see Zubin & Köpcke, 2009). Where do such groups come from? Could they be arbitrary?

Zubin and Köpcke (2009) attribute those generalizations to semantic fields. When those fields become productive, they reach the status similar to a construction. Constructions are in turn capable of transferring grammatical gender to noun phrases. Therefore, this transfer of gender is part of the construction and hence lexicon, while at the same time it is conceptually triggered and thus it is part of pragmatics as well. Interestingly, they apply the term natural gender to inanimates as well, and their pragmatic projection hypothesis applies to both animates and inanimates (Zubin & Köpcke, 2009, p. 249). For example, the noun car has natural masculine gender.
Figure 2 shows that the German noun *das Auto* “car” in general use belongs to neuter gender. However, according to the authors and the example from Figure 3, when it is a particular model in one’s mind, he uses masculine gender. “The apparent source of masc-gender for the pronoun is a pragmatic projection associated with the concept ‘car’, evoked by the picture of a car in the ad.” (Zubin & Köpcke, 2009, p. 242)

Finally, gender is a category that tackles morphology, syntax, semantics and cognition. Therefore, it deserves our attention. Moreover, it is the category easily acquired by children. In the Section 6.1.1 (this thesis), we have seen that diminutives play an important role in making gender less obscure and more simple, and hence aids in acquiring it faster. As Moskovljević-Popović (personal conversation, October 30, 2014) said, in Serbian children make very few errors, e.g. the noun *kost* “bone,” with an atypical feminine ending, is sometimes misused as if it belonged to the masculine gender. Of course, there are a lot of clues for gender in Serbian and Croatian. Gender is hidden in both verb and noun phrases in a sentence.

As this chapter showed, the distinction should be made between the synchronic and the diachronic view to the system of gender. On the synchronic level we should perceive and divide two different perspectives of gender – natural and grammatical. Natural gender becomes important only in the cases of agreement with verbs. For example, it is grammatically correct to say both *kolovođ-e su pobjegl-e* and *kolovođ-e su pobjegl-i* “ringleaders ran away,” meaning that we can concord according to both the grammatical and the semantic rule, respectfully (Barić et al., 1997, p. 424).

Diachronic level, on the contrary, offers us a better understanding of the system of gender in general. It enables us to explore the relationship between natural and grammatical gender. We can inspect motivations for gender and its relationship to other categories, such as case for example (see Toyota & Vlasa Florea, 2009).

7. Conclusion

A noun inflection system represents an important segment of language and cognition. Moreover, a profound understanding of the Croatian noun inflection system enabled further analysis of gender assignment systems. Namely, in section 2, two noun inflection classification systems were observed – the traditional one and the Natural Morphology account on Serbo-Croatian noun inflection classification. The traditional one offers us a rich description of the noun inflection system, with all the exception rules defined and explained. No matter how important descriptive
information it contains, due to its vagueness and the frequent usage of words such as “sometimes,” “usually,” “most,” and similar, it cannot be easily applied within the computer analysis (cf. Tadić, 1994). The Natural Morphology approach to noun classification is based on the productive and the unproductive morphological rules (which further leads to forming different macroclasses, classes, subclasses and microclasses). Therefore, Natural Morphology recognizes the dynamics of morphology and has an aim to explore its breathing. The main benefits of the Natural Morphology account for this thesis were the pieces of information related to the productiveness of the patterns. For instance, the rule that –o and -e are productive masculine endings enabled the CroGIM to have a general rule for assigning masculine gender to such nouns. The importance of such a rule is enormous, particularly because in the traditional grammar it is often stated that most nouns ending in –o/-e are neuter (see e.g. Barić et al., 1997). Even though there are more neuter than masculine nouns ending in –o/-e, only the masculine ones are productive.

After observing the Croatian noun inflection classes in section 2, the focus of the thesis shifted to defining gender and gender assignment approaches in section 3. In section 3, the theoretical approaches to gender assignment that are related to computer analysis were presented. In particular, the approach incorporated in the Corpus and the Network morphology approach (as defined in Corbett & Fraser, 1993). The relevance of the first of these two approaches reflects the fact that the Croatian language is being investigated within this thesis. The second one, the Network morphology approach served for the making of the CroGIM. Namely, defaults, generalizations, as well as rule overriding are important notions within both the CroGIM and the Network Morphology approach. However, the main differences lie in handling exceptions and in the perception of productivity. In other words, exceptions in the CroGIM are listed before the generalizations are made, while within the Corbett and Fraser (1993) exceptions are defined separately after the generalization is made. Also, Corbett and Fraser (1993) do not benefit from productivity, whereas CroGIM does. With regards to gender assignment, Corbet and Fraser (1993) assign gender from the relationship between nouns and adjectives only in case it is not clear from the declension class. CroGIM, on the other hand, uses the living information from the Corpus, in order to connect to reality and detect whether the gender of a noun has changed. This connection enables us to track the history and observe how the language breathes.

As for the CroGIM performance, it opened a lot of questions regarding the definition of gender as well as the problems of gender assignment. If we would like to take a look at all the rules in the CroGIM, it could be easily seen that exceptions to general rules are often words with marked and specific use, such as diminutives, augmentatives, hypocoristics, etc.

To conclude, a lot of questions remained open in section 5. One such question is of the issue of relations between proper and common nouns, or names that could refer to both females and males. The fascinating thing is that gender often seems to be one of the clearest categories in a language to native speakers, e.g. in Croatian, nouns are said to belong to masculine, feminine or neuter gender (i.e. Barić et al., 1997). Still, it is only after engaging into building a computational model capable of assigning gender, that one is able to conceive the great dimness of the category of gender.
8. Bibliography


MacWhinney, B. (2010). A tale of two paradigms. In M. Kail & M. Hickmann (Eds.), Language acquisition across linguistic and cognitive systems. Amsterdam: Benjamins


48

9. Appendix

A. The Python code

class gen(object):
    semantic = ''
    singular = ''
    plural = ''

def outprint(self):
    answer = ''
    if len(self.singular) > 0:
        if(self.singular == self.plural == self.semantic):
            return self.singular
        else:
            answer += 'singular: ' + self.singular
    if len(self.plural) > 0:
        if(len(answer) > 0):
            answer += ', '
        answer += 'plural: ' + self.plural
    if len(self.semantic) > 0:
        if(len(answer) > 0):
            answer += ', '
        answer += 'semantic: ' + self.semantic
    return answer
def setAll(self,newGender):
    self.singular = newGender
    self.plural = newGender
    self.semantic = newGender

class Objekat(object):

    def __init__(self):
        self.gender = gen()
        self.value = ''
    
    def outprint(self):
        return 'gender || ' + self.gender.outprint()

class genderCodeList:
    masculine = 'masculine'
    feminine = 'feminine'
    neuter = 'neuter'

inputedValue = (str (input ()))
X = Objekat()
g = genderCodeList()
X.value = inputedValue

def endx (i):
    if i != '):
        return i[-1]

    X.gender.setAll(g.neuter)
    X.gender.setAll(g.masculine)
    X.gender.setAll(g.masculine)
    X.gender.setAll(g.neuter)
elif X.value in ['pista','lista','rang-lista','bista','top-lista','nevista','cista','transferlista','linklista','glista','cast-lista']:
    X.gender.setAll(g.feminine)
elif X.value in ['kino', 'dinamo', 'torzo', 'libreto', 'moto', 'molo', 'veto', 'salto', 'korzo', 'geto', 'torero', 'mikado']:
    X.gender.singular = g.masculine
    X.gender.plural = g.neuter
elif X.value in ['doba']:
    X.gender.setAll(g.neuter)
else:
    elif X.value in ['misao', 'pomisao', 'izrastao', 'so', 'mati']:
        X.gender.setAll(g.feminine)
    elif X.value in ['dečače', 'momče']:
        X.gender.setAll(g.neuter)
        X.gender.semantic = g.masculine
    elif X.value in ['devojče']:
        X.gender.setAll(g.neuter)
        X.gender.semantic = g.feminine
        X.gender.setAll(g.masculine)
    elif X.value in ['tata', 'vojvoda', 'zanatlija', 'papa', 'zavičajlija', 'delija', 'novajlija', 'poslovoda', 'starešina', 'sudija', 'sluga', 'kalf', 'djelovoda']:
        X.gender.setAll(g.masculine)
X.gender.setAll(g.feminine)

X.gender.setAll(g.feminine)

elif X.value in ['ľelja', 'povelja', 'nedelja', 'znatiľelja', 'kekcelja', 'stelja', 'nedjelja', 'fotelja'] or X.value in ['Zdelja', 'Šv.Nedjelja', 'Ljelja', 'Lelja']:
X.gender.setAll(g.feminine)

X.gender.setAll(g.feminine)

X.gender.setAll(g.feminine)
    X.gender.setAll(g.masculine)

elif len(X.value) > 4 and X.value[-5:] in ['alija', 'elija', 'ilija', 'oliya', 'uliya']:
    X.gender.setAll(g.feminine)

elif X.value[-4:] == 'ište':
    X.gender.setAll(g.neuter)

elif X.value[-4:] in ['džija', 'lija'] and not X.value.istitle():
    X.gender.setAll(g.masculine)

elif X.value[-2:] == 'ša':
    X.gender.setAll(g.masculine)

elif X.value[-2:] == 'vo':
    X.gender.setAll(g.neuter)

elif X.value[-2:] == 'će':
    X.gender.setAll(g.neuter)

elif X.value[-2:] == 'lo' and not X.value.istitle() and not X.value in ['a-raphael-tiepolo', 'diavolo', 'discepolo', 'escenabolo', 'karlo', 'mullo', 'mulo', 'polo', 'rolo', 'vaterpolo'] or X.value in ['Etnoselo']:
    X.gender.setAll(g.neuter)

elif X.value[-2:] == 'lo' and X.value.istitle():
    X.gender.setAll(g.masculine)

elif X.value[-2:] == 'ce':
    X.gender.setAll(g.neuter)

elif X.value[-2:] == 'če' and not X.value.istitle():
    X.gender.setAll(g.neuter)

elif X.value[-3:] == 'ina' and X.value in ['budaletina', 'detektivčina', 'dimničarina', 'djedetina', 'djedurina', 'dječačina', 'gazdetina', 'generalina', 'glaveština', 'ljudina', 'muškarčina']:
    X.gender.setAll(g.feminine)
X.gender.semantic = g.masculine

elif X.value[-3:] == 'ina':
    X.gender.setAll(g.feminine)

elif X.value[-3:] in ['nje','lje','ţĎe']:
    X.gender.setAll(g.neuter)

elif X.value[-2:] == 'je':
    X.gender.setAll(g.neuter)

elif X.value[-3:] == 'ost' and X.value not in ['gost', 'most', 'oprost', 'post']:
    X.gender.setAll(g.feminine)

elif not X.value[-4:] in ['onja','elja'] and X.value[-1] == 'a':
    X.gender.setAll(g.feminine)

elif X.value[-4:] == 'elja':
    X.gender.setAll(g.masculine)

elif X.value[-4:] == 'onja':
    X.gender.setAll(g.masculine)

elif X.value[-1] in ['o', 'e', 'i', 'u', 'y']:
    X.gender.setAll(g.masculine)

print(X.outprint())

B. The JavaScript code

var vowels = "aeiou";

var g = {
    masculine : 'm',
    neuter : 'n',
    feminine : 'f'
};

exports.determineGenderFromAdjective = function(adjective){
var adjectiveEnding = adjective[adjective.length - 1];

var gender;

if(adjectiveEnding === 'i' || vowels.indexOf(adjectiveEnding) === -1){
    gender = 'm';
}
else if(adjectiveEnding === 'o' || adjectiveEnding === 'e'){
    gender = 'n';
}
else if(adjectiveEnding === 'a'){
    gender = 'f';
}

return gender;

}

exports.determineGenderFromNoun = function(noun){
    var X = {
        value : noun,
        gender : {
            semantic : ",",
            singular : "",
            plural : "",
            setAll : function(value){
                this.semantic = this.singular = this.plural = value;
            }
        }
    }
}


X.gender.setAll(g.neuter);

}


['genije', 'foaje', 'denije', 'amfilohije', 'atelje', 'dosije', 'dosje'].indexOf(X.value) !== -1)

X.gender.setAll(g.masculine);

}


X.gender.setAll(g.masculine);

}


X.gender.setAll(g.masculine);

}

    X.gender.setAll(g.neuter);
}
else if (['pista','lista','rang-lista','bista','top-lista','nevista','cista','transfer-lista','linklista','glista','cast-lista'].indexOf(X.value) !== -1){
    X.gender.setAll(g.feminine);
}
else if (['kino', 'dinamo', 'torzo', 'libreto', 'moto', 'molo', 'veto', 'salto', 'korzo', 'geto', 'torero', 'mikado'].indexOf(X.value) !== -1)
    X.gender.singular = g.masculine;
    X.gender.plural = g.neuter;
}
else if (['doba'].indexOf(X.value) !== -1)
    X.gender.setAll(g.neuter);
}
else if (['misao', 'pomisao','izrastao', 'so', 'mati'].indexOf(X.value) !== -1){
    X.gender.setAll(g.feminine);
}
else if (['dečače', 'momče'].indexOf(X.value) !== -1)
    X.gender.setAll(g.neuter);
X.gender.semantic(g.masculine);
}
else if (['devojče'].indexOf(X.value) !== -1)
    X.gender.setAll(g.neuter);
X.gender.semantic = g.feminine;
    X.gender.setAll(g.masculine);
}

else if (['tata', 'vojvoda', 'zanatlija', 'papa', 'zavičajlija', 'delija', 'novajlija', 'poslovoĎa', 'starešina', 'sudija', 'sluga', 'kalfa', 'djelovoda'].indexOf(X.value) !== -1){
    X.gender.setAll(g.masculine);
}

    X.gender.setAll(g.feminine);
    X.gender.setAll(g.feminine);
}
else if (['kiša', 'piša', 'polija', 'gerija', 'niša', 'čaša', 'kaša', 'snaša', 'šalša'].indexOf(X.value) !== -1 || ['Urša', 'Maša', 'Deša', 'Ajša'].indexOf(X.value) !== -1){
    X.gender.setAll(g.feminine);
} else if (['želja', 'povelja', 'nedelja', 'znatiželja', 'kece', 'stelja', 'nedelja', 'fotelja'].indexOf(X.value) !== -1 || ['Zdelja', 'Sv. Nedelja', 'Ljelja', 'Lela'].indexOf(X.value) !== -1){
    X.gender.setAll(g.feminine);
}
    X.gender.setAll(g.feminine);
}
    X.gender.setAll(g.feminine);
}
    X.gender.setAll(g.feminine);
}
}
X.gender.setAll(g.masculine);
}
else if (X.value.length > 4 && ['alija', 'elija', 'ilija', 'oliya', 'ulija'].indexOf(X.value.substr(X.value.length - 5)) !== -1)
X.gender.setAll(g.feminine);
}
else if (X.value.substr(X.value.length - 4) === 'ište')
X.gender.setAll(g.neuter);
}
else if (['džija', 'ljija'].indexOf(X.value.substr(X.value.length - 4)) !== -1 && (X.value[0].toUpperCase() !== X.value[0]))
X.gender.setAll(g.masculine);
}
else if (X.value.substr(X.value.length - 2) === 'ša')
X.gender.setAll(g.masculine);
}
else if (X.value.substr(X.value.length - 2) === 'vo')
X.gender.setAll(g.neuter);
}
else if (X.value.substr(X.value.length - 2) === 'će')
X.gender.setAll(g.neuter);
}
else if (X.value.substr(X.value.length - 2) === 'lo' && (X.value[0].toUpperCase() !== X.value[0]) && ['a-raphael-tiepolo', 'diavolo', 'discepolo', 'escenabolo', 'karlo', 'mullo', 'mulo', 'polo', 'rolo', 'vaterpolo'].indexOf(X.value) !== -1 || ['Etnoselo'].indexOf(X.value) !== -1)
X.gender.setAll(g.neuter);
}
else if (X.value.substr(X.value.length - 2) === 'lo' && X.value[0].toUpperCase() == X.value[0])

X.gender.setAll(g.masculine);
}
else if (X.value.substr(X.value.length - 2) === 'će'){
X.gender.setAll(g.neuter);
}
else if (X.value.substr(X.value.length - 2) === 'će' && X.value[0].toUpperCase() !== X.value[0]){
X.gender.setAll(g.neuter);
}
else if (X.value.substr(X.value.length - 3) === 'ina' && ['budalina', 'detektičina', 'dimničina', 'djedetina', 'dječurina', 'dječačina', 'gazdetina', 'generalina', 'glaveština', 'ljudina', 'muškarčina'].indexOf(X.value) !== -1){
X.gender.setAll(g.feminine);
X.gender.semantic = g.masculine;
}
else if (X.value.substr(X.value.length - 3) === 'ina'){
X.gender.setAll(g.feminine);
}
else if (['nje', 'lje', 'žde'].indexOf(X.value.substr(X.value.length - 3)) !== -1){
X.gender.setAll(g.neuter);
}
else if (X.value.substr(X.value.length - 2) === 'je'){
X.gender.setAll(g.neuter);
}
else if (X.value.substr(X.value.length - 3) === 'ost' && ['gost', 'most', 'oprost', 'post'].indexOf(X.value) === -1){
X.gender.setAll(g.feminine);
}
else if (['onja','elja'].indexOf(X.value.substr(X.value.length-4)) === -1)
  X.gender.setAll(g.feminine);
}
else if (X.value.substr(X.value.length - 4) === 'elja'){
  X.gender.setAll(g.masculine);
}
else if (X.value.substr(X.value.length - 4) === 'onja'){
  X.gender.setAll(g.masculine);
}
else if (['o', 'e', 'i', 'u', 'y'].indexOf(X.value[X.value.length-1]) !== -1)
  X.gender.setAll(g.masculine);
else if (['b', 'v', 'g', 'd', 'Ď', 'z', 'ţ', 'j', 'k', 'l', 'lj', 'p', 'r', 't', 'ć', 'f', 'h', 'c', 'č', 'dţ', 'š'].indexOf(X.value[X.value.length-1]) !== -1)
  X.gender.setAll(g.masculine);
}
return X;