Recognizing Diminutive and Augmentative Croatian Nouns

Kristina Kocijan[1], Marijana Janjić[1], Sara Librenjak[1]

[1]Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia

{krkocijan, marijanji, slibrenj}@ffzg.hr

Abstract. In this paper, the authors present NooJ morphological grammars for recognizing Croatian diminutive and augmentative nouns for those common nouns that already exist in the Croatian NooJ dictionary. The purpose of this project is twofold. The first one is to recognize both diminutive and augmentative forms of each noun existing in our dictionary (over 20 000 common nouns) if such a form occurs in a text. The second purpose is to determine types of texts in which these words appear the most (or if they even appear) which is the reason why we divided our corpus in two thematic categories (children literature, novels). The results of our algorithm are high on both types of text [overall P=0.82; R=0.80; f-measure=0.81]. Although NooJ dictionary allows direct entrance of such derivations as an attribute-value description of a main noun, we have opted for the second option, i.e. writing a morphological grammar that will recognize the needed form. In this way, we are saving the space and time needed to add all the existing forms to the noun's dictionary.

Keywords: augmentatives; diminutives; Croatian; morphology; nouns; NooJ.

1 Introduction

Diminutives and augmentatives are a fruitful field of study as can be seen from the number of approaches taken up by authors in their research across languages: phonetics [10], morphology [5], [6], [8], semantic aspects [7], [13], [15] to name but a few. The most relevant work for this paper is the research in Croatian language, as our paper aims to add some new information and insight in diminutives and augmentatives in Croatian. The search for such research did not give us many results, but it was diverse and rich in information that helped us in our preliminary steps.

adfa, p. 1, 2011.

The authors of Croatian grammars and older generation of researchers [1, 2] focused mainly on the morphological issues in diminutives and augmentatives in general, offering a short overview of the dominant morphological patterns. On the other hand, another approach is taken in some more recent research papers. Thus, analysis of diminutive Croatian verbs was discussed in details in [9], while [4] and [16] provide a more semantic and pragmatic view to the topic of Croatian diminutives. Some comparisons between diminutive and/or augmentative suffixes of Croatian and other languages are found in [15], [16], [19] among others.

The overview of the data shows that the definitions of diminutives and augmentatives are an issue on its own. Bosanac et al. [4] speculate that the lack of precise definitions in grammar might be caused by the lack of field research on the use of such lexemes. Thus, their paper points out to a great need for pragmatic scholarly research that either involves questionnaires, as is the case of research conducted by [4], or the corpus based research, as is the case in this paper.

We believe that this paper will be a small contribution to a study oriented towards the language usage with its two aims. The first one is to recognize both diminutive and augmentative forms of each noun existing in the Croatian NooJ dictionary (over 20 000 common nouns [20]) if such a form occurs in a text. At this point verbal and adjective forms are not taken into account as that would make the task very complex. Hence, we agreed to focus on a segment of a puzzle in order to test the methodology.

The reason for the variation in the corpus is the second question that concerns us: in which types of texts diminutive and augmentative nouns mostly appear. The reasons behind the combination of corpus sources is not just that the earlier scholarly work was mostly theoretical, but also that it did not always clearly state how certain results and statistics were gained. Thus, [1, 2] discuss the productivity of particular suffixes in nominal diminutives but does not offer information on the corpus this statistics is based on. In [1] it remains unclear why the year 1860 was relevant for the author's research on contemporary Croatian language, or what is the information on some of the most productive suffixes based on, as the author offers an explanation of ‘general language acquaintance’ as a relevant one. Such issues point out the importance of critical analysis of previous research in order to offer some new insight in the field of (Croatian) diminutives and augmentatives in general.

The first step as a step towards attaining the new insight is to build morphological grammars that will be able to recognize diminutive and augmentative occurrences regardless the number and the case of a noun. Similar work has already been reported for Serbian [11] and Portuguese [14]. Once built, the grammars would be able to recognize and classify the newly found nouns either as a diminutive or augmentative form of a noun existing in the NooJ dictionary [20]. For example, if there is a noun *kuća* (house) in the dictionary, the grammar will also recognize its diminutive form *kućica* (en. small house) and mark it as <N+DIM+MAIN=kuća +Case+Gender+Number>. The augmentative form *kućerina* (en. big house) will be marked after the same pattern as <N+AUG+MAIN=kuća+Case+Gender+Number>. The only prerequisite is that the main noun exists in the NooJ dictionary. In this paper, we will present how we built the morphological grammar, what issues have we encountered and how we propose to solve them. The conclusions that we offer are based on the experience we gained from building grammars and corpus analysis. An interested reader should however bear in mind that this is only the first phase of the research and hence, not all answers are available at the moment.

The remaining of this paper is structured in the following manner. We will start with providing some theoretical approaches that exist for Croatian diminutive and augmentative nouns and then will turn to the digital dictionary of nouns that we used as the basic platform for our grammars. We will proceed with an insight into the corpus we have used for this research that will be followed by the description of our grammars for detecting and annotating derivational forms. At the end of the paper, we will give the results we have collected so far and will provide some final thoughts on our results and the usability of our approach.

2 The Theory behind Diminutives and Augmentatives in Croatian Language

When we talk about diminutives and augmentatives, we are talking about nouns that are much smaller (diminutives) or much bigger (augmentatives) in size or value or intensity than the average [4].

Contexts in which diminutive forms appear are mostly observed in **child directed speech** (*Daj mi lopticu*[1]*!* – en: Give me [little] ball), **teasing** (*Ti si moja slatka loptica*[2]*.* – en. You are my sweet [little] ball.), **affectionate utterance** (*Da vidimo taj vaš trbuščić*[3]*.* – en. Let's see that [little] belly of yours.) and in **jargon** (*Bokić*[4]. – en. [Little] hello.). Augmentative forms have been detected when **making fun of somebody or something** (*Koja glavurina.* – en. What a [huge] head.), when **insulting someone** – (*Ona je kravetina*. – en. She is a [huge] cow.; *On je konjina!* – en. He is a [huge] horse!), when **exaggerating** (*Koja zgradurina*. – en. What a [huge] building.) or for **exaltation of a positive characteristic** (*On je prava ljudina*[5]. – en. He is a [very big] person.). However, some times, depending on the context, some diminutive nouns may take augmentative meaning (ex. *psihić*[6]) and vice versa (ex. *glavonja*[7]) [16]. More on the context and the meanings of Croatian diminutives may be found in [4].

The list of suffixes used to build diminutive and augmentative forms is a close set. Some suffixes are more productive and are used for several hundred of nouns while some are used for only one or several nouns. The lists are gender dependent i.e. some suffixes are characteristic for only masculine (*-ić | -čić | -(a)k | -eč(a)k | -ič(a)k*), or only feminine (*-ca | -ica | -čica*), or only neutral nouns (*-ce | -ance | -ašce | -ence | -ešce*). Still, there are suffixes like –

[1] Expression is used regardless the size of a ball.
[2] Expression used for an overweight person.
[3] Expression used for a pregnant lady regardless the size of her belly.
[4] Greeting characteristic for the area of City of Zagreb.
[5] Expression used for a person of a big heart.
[6] Diminutive form of a noun 'psychiatrist' when talking about a great psychiatrist.
[7] Augmentative form of a noun 'head' when talking about someone who is always proposing some not so good ideas.

eljak and *–uljak* for diminutives that may be used for any gender nouns. Suffixes for augmentative nouns are *-ina | -čina | -etina | -urina| -erina | -ešina | -ura | -urda | -ušina | -uština | –eskara | -uskara* regardless the noun's gender. As in the case of some diminutives, a gender can be shifted. Consequently, masculine nouns that take suffixes *–ina, -čina, -etina, -urina, -usina,* or *–čuga* may either remain masculine nouns or may become feminine nouns.

In Croatian language we can produce diminutive and/or augmentative forms for the large number of nouns (both common and proper). In the course of work on NooJ grammar we have realized, however, that certain limits do exist and that there are nouns which do not produce diminutive and/or augmentative forms. Such are, for example, abstract nouns like *milosrđe (*en. mercy*)* or collective nouns like *cvijeće (*en. flowers*), suđe (*en. dishes*), unučad (*en. grandchildren*), otočje (*en. islands) etc. which do not form either diminutives or augmentatives.

Most of the common neutral nouns that end in *–nje* are considered to be verbal nouns (i.e. they were built from verbs) and they do not make diminutive or augmentative forms either. The same applies to some other neutral nouns ending in *–nje* like *bezakonje (*en. lawlessness) or *trnje (*en. thorns*),* but there are some that are exceptions to this rule like *janje (*en. lamb*).*Majority of common neutral nouns ending in *–ce* (*lice* – en. face*), -če* (*unuče*– en. grandchild*), -će* (*proljeće*– en. spring*), -đe* (*suđe*– en. dishes*), -je*[8] (*otočje* – en. islands), *-lje* (*bilje*– en. plants), *-šte (stubište*– en. stairway*), -vo (pecivo* – en. pastry*), -stvo (sudstvo*– en. judiciary*), -štvo (divljaštvo*– en. savagery*), -ost (jednostavnost* – en. simplicity) do not have their diminutive or augmentative forms either.

At this stage we also became aware of the nouns which appear to be diminutives but they have ceased to be perceived that way by Croatian speakers, or they co-exist in both diminutive and non-diminutive meanings. Such is the case with the following examples: *vreća* (en.bag) – *vrećica* (en. little bag but

[8] However, *jaje* (en. egg) is an exception.

also grocery bag)*,* *vrt* (en. garden) – *vrtić* (en. kindergarten), *ploča* (en. panel) – *pločica* (en. little panel but also bathroom tile).

Next to the nouns that do not have diminutive forms, Babić's work [1, 2, 3] also made us aware of the nouns that produce several diminutive forms but with different meanings (ex. *kuća* (en. house): *kućica* – little house and *kućerak* – little poor house [1]). Such words present a specific semantic problem that falls outside the scope of this paper and thus will not be discussed here in more details.

There is also a class of nouns that appears to bear one of the augmentative morphemes, but actually only has a visual resemblance to augmentatives while their phonological elements differ [12]. For example, noun *pile* (en. chicken) when added suffixes *–ina* (which is an augmentative suffixes) builds a noun *piletina* with the meaning 'chicken meat' and not 'a [huge] chicken'. Another issue related to diminutives is their relation with hypocorisms in Croatian language, while augmentatives are also interrelated with pejoratives.

In the next section, we will give a short account of the Croatian nouns in NooJ dictionary before we turn to the detailed description of morphological grammars for detection of diminutives and augmentatives, that we have built for this project.

3 Dictionary of Croatian Nouns

Presently, there are 20 350 common nouns in Croatian NooJ Dictionary, of which 8 416 are feminine, 6 387 are masculine and 5 547 are of neutral gender [20]. However, 4 423 neutral nouns do not form either diminutive or augmentative forms. We can recognize this type of nouns by their type and gender information in addition to their endings. We use this information to seclude them from our morphological grammar recognition.

Since NooJ [18] allows the derivational paradigms to be added directly to the dictionary, one of the possible approaches to recognize diminutive and augmentative forms is to add +DRV attribute to each of the existing nouns for each derivation that is possible. Thus, while some of the nouns have only one derivation for either diminutive or augmentative form, some have both derivations, some neither and some may have several diminutive and/or augmentative derivations. Examples for the last type are found in [2] *kamen -> kamenčina, kamenčuga, kamenina*[9] (en. stone -> big stone), *noga -> nogetina, nožetina* (en. leg -> big leg)[10]. There are also nouns that share the diminutive or augmentative form, ex. *repetina* is both '*veliki rep*' (en. big tale) and '*velika repa*' (en. big beet). This direct definition of derivations inside the dictionary would definitely result in 100% precision, and a very high recall, but it may not detect any new entries in the text that language, as a living thing, makes possible. This is the main reason that we have opted to pursue another path i.e. building the morphological grammar for detection of diminutive and augmentative nouns in Croatian. The other reason falls in the domain of too many person hours needed to add each derivation manually directly to the dictionary.

4 Corpus

We have divided our corpus into 2 main categories in the following manner: CAT1 consists of children stories, and CAT2 of novels, mainly by Croatian authors. Bosanac et al. [4] have found six basic semantic and pragmatic features of diminutives in Croatian: the object in question is in diminutive form because it is considered small, seen with affection, considered negative (pejorative meaning), large or neutral. Additionally, the meaning can be contextualized or lexicalized. "Small" and "affectionate" were most prominent semantic connotations of diminutives. According to [7], diminutives are used as a politeness and softening device in discourse, so we can hypothesize that they would be more common in texts containing dialogues, such as novels.

[9] Some derivations (like *kamenina*) are only rarely used and found mainly in literal texts.

[10] Some of the examples are, however, limited to individual usage. Thus the example *kamenina* which Babić uses has been used by the literary writer Božić [2], while many average Croatian speakers would find it unusual and eccentric.

Considering this, we have chosen two stylistically different corpora in order to analyze differences in the usage of stylistically specific words, i.e. diminutives and augmentatives, in different environments. We assumed that novels and stories directed towards children contain a larger amount of diminutives and augmentatives then texts directed towards adult audience, unless specifically needed in a context. Children corpus would most likely contain more words with "small" and "affectionate" meaning while novels may contain a number of diminutives and augmentatives for their poetic and stylistic meanings.

Table 1. represents the size of each of the two corpora. We have assumed that children's novels (CAT1) would contain more diminutives then regular novels (CAT2). This has been proven true, as we have found 7 times more diminutives by their relative frequency in CAT1 (0.96%) then in CAT2 (0.13%). However, although we expected to find significantly more augmentatives in CAT1 as well, their number was about the same in both corpora.

Table 1. The size and frequency of diminutives and augmentatives in the corpora constructed for testing purposes

Category	Number of tokens	Number of diminutives		Number of augmentatives	
CAT1	240 616	2 312	0.96%	114	0.047%
CAT2	504 876	696	0.14%	59	0.032%

For the purpose of this research, we have chosen to compare the results obtained manually (human annotators) and automatically (NooJ morphological grammar) in these two categories (CAT1 and CAT2). We present our findings in the Results section.

5 Recognizing Diminutive Nouns

The morphological grammar for recognizing diminutive Croatian nouns consists of four main graphs depending on the gender related endings. The first graph recognizes diminutives characteristic for the feminine nouns (including diminutive suffixes: *-ca, -ica, -čica*), the second graph for masculine nouns (*-ac, -ak, -čić, -ečak, -ičak*) (Fig. 1) and the third graph for neutral nouns (*-ce, -ance, -ašce, -ence, -ešce*).The fourth graph is for recognizing diminutives build with endings that are not related to gender and as such includes masculine, feminine and neutral nouns (*-ić, -če, -eljak, -erak, -uljak*). Still, the situation is not as clear as it may appear since there are some suffixes that are used in different gender nouns [1]. This is not only a characteristic for Croatian diminutives but it is also observed in other languages like Romanian [7], Slovene [17] or Slovak and Czech [15]. Babić [1] gives a list of 25 diminutive suffixes in total, three of which are highly productive (*-ica, -ić, -čić*), one is partially productive (*-če*) while others are considered to be weakly or non-productive suffixes.

The main logic behind each of the graphs is to take the unknown word, divide it into three sections: N = main noun, S = diminutive particle, and case ending; and check the following:

a) does the section N exists in the main dictionary as the common noun in singular Nominative (in the gender defined by the subgraph type: feminine, masculine, neutral or any gender);
b) what is the diminutive particle S;
c) what is the case ending.

Finally, each recognized word is annotated as a common noun N, of type=*umanj* (short for diminutive in Croatian), and the case and number depending on the case ending defined in ***c,*** with the main noun N as its superlemma. Connecting diminutives and augmentatives with their superlemma will allow us to find all the forms of a noun *kuća* (en. house) whether it is written in the regular form or as a diminutive and augmentative form.

The graph depicting the recognition of diminutives for masculine nouns (Fig. 1) is given here as an example since the same logic is followed in the remaining graphs for diminutives. This graph consist of more than one path (as one might expect). This is due to the fact that Croatian morphology requires some phonological alternations when particular phonemes are in the same environment, such as changing 'b' for 'p' (upper path in Fig. 1) or shortening the 'ije' to 'je' (lower paths in Fig. 1). So for example, deminutive for the noun *'cv**ije**t'* (en. flower) is *'cv**je**tić'* (en. little flower).

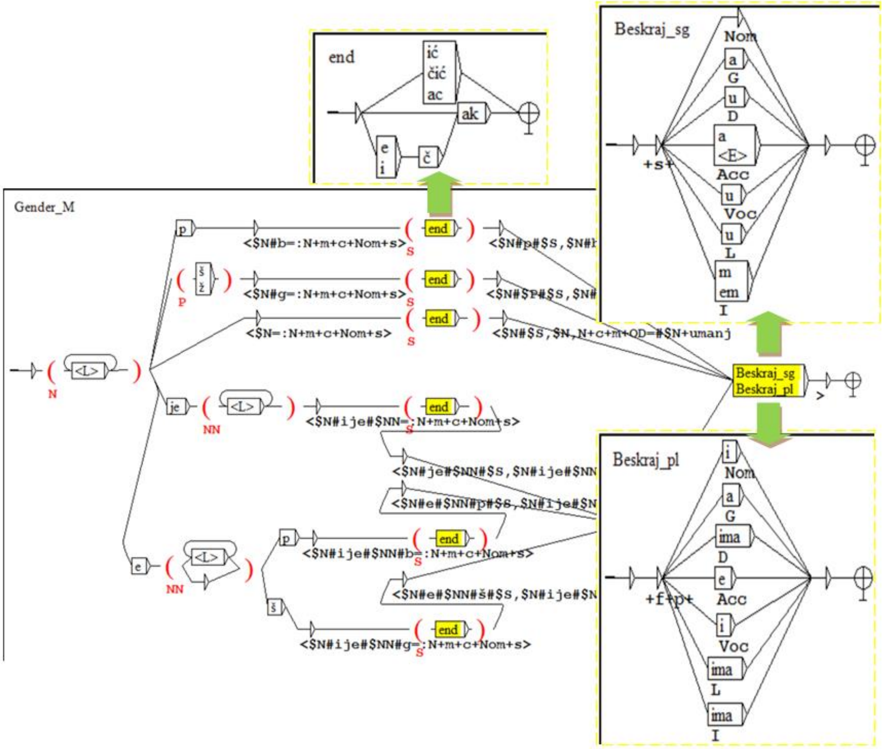

Fig. 1. Section of a morphological grammar recognizing diminutives characteristic of masculine nouns

Such alternations required separate approach in the grammar which resulted with some extra paths in the graph. Due to these and similar alternations, the section of a word marked as N is not always a root noun in its full form. Thus, the missing, or alternated phonemes had to be glued back to the N before we can check if the root noun exists in the dictionary. This has been done with a NooJ special character '#' that lets us concatanate more strings into one. Thus, the expression '**$N#ije'** adds the string 'ije' directly to whatever string is found in the variable $N. As it will be seen in the following section, we have utilized this feature in the grammar for recognizing augmentative nouns as well, since we were dealing with the same situations in detecting the root noun.

6 Recognizing Augmentative Nouns

In [16] there are 23 and in [2] even 30 different suffixes that are used for building augmentative noun forms in Croatian. We have used them all in our grammar (Fig. 2) although only six of these are considered to be very productive: *-ina, -čina, -etina, -urina, -jurina* and *–usina*, while others, like *–enda, endra, -erda, -(j)urda, -urenda* or *–čuga* are used for either only one or just a few nouns.

The main graph is divided into three possible paths, depending on the gender of the noun whose augmentative form we are recognizing. All three paths merge again into a single subgraph that describes the case endings of augmentative nouns. Inside the gender related subgraph we have described all the possible changes that may happen to the main noun before the augmentative suffix is added. Thus, if we follow the first path in the subgraph **Gender_F_**, we need to check if the string recognized prior to the augmentative suffix, presenting the main noun, exists in the dictionary as a common singular noun in feminine gender and Nominative case.

However, this is not possible for those nouns that loose or change their ending before the augmentative suffix is added. To illustrate this phenomenon, let us take a look at the following example. The common singular feminine noun in Nominative case *kabani**ca*** (en. raincoat) has an augmentative form *kabani+**četina*** (en. large raincoat). At first we observe that the ending –**ca** from *kabani**ca*** does not appear when the suffix is added. From Croatian Grammars we know that this process is called palatalization and it is described as: *kabanic-a → kabanic + etina → [c + e → č + e] → kabaničetina* (i.e. delete the last letter ‘a’ and add the suffix ‘*etina’*; if ‘c’ is found before the ‘e’ from the suffix ‘*etina*’, change the ‘c’ into ‘č’). In order to simulate this process, we need to append ‘ca’ at the end of a recognized string (‘*kabani’*) that is placed in a variable $N but only where ‘č+etina’ follows the variable $N. Again, by using NooJ notation described in the previous section **<$N#ca =: N+c+f+Nom+s>** we are able to check if a feminine common noun ‘*kabanica’* exists in our dictionary. Only if a noun from the variable $N#ending exists in our dictionary, we are able to proceed along our path which takes us to the content of variable $S. This variable holds an entire new subgraph *aug_F* that holds all the possible endings used for building augmentatives of feminine nouns.

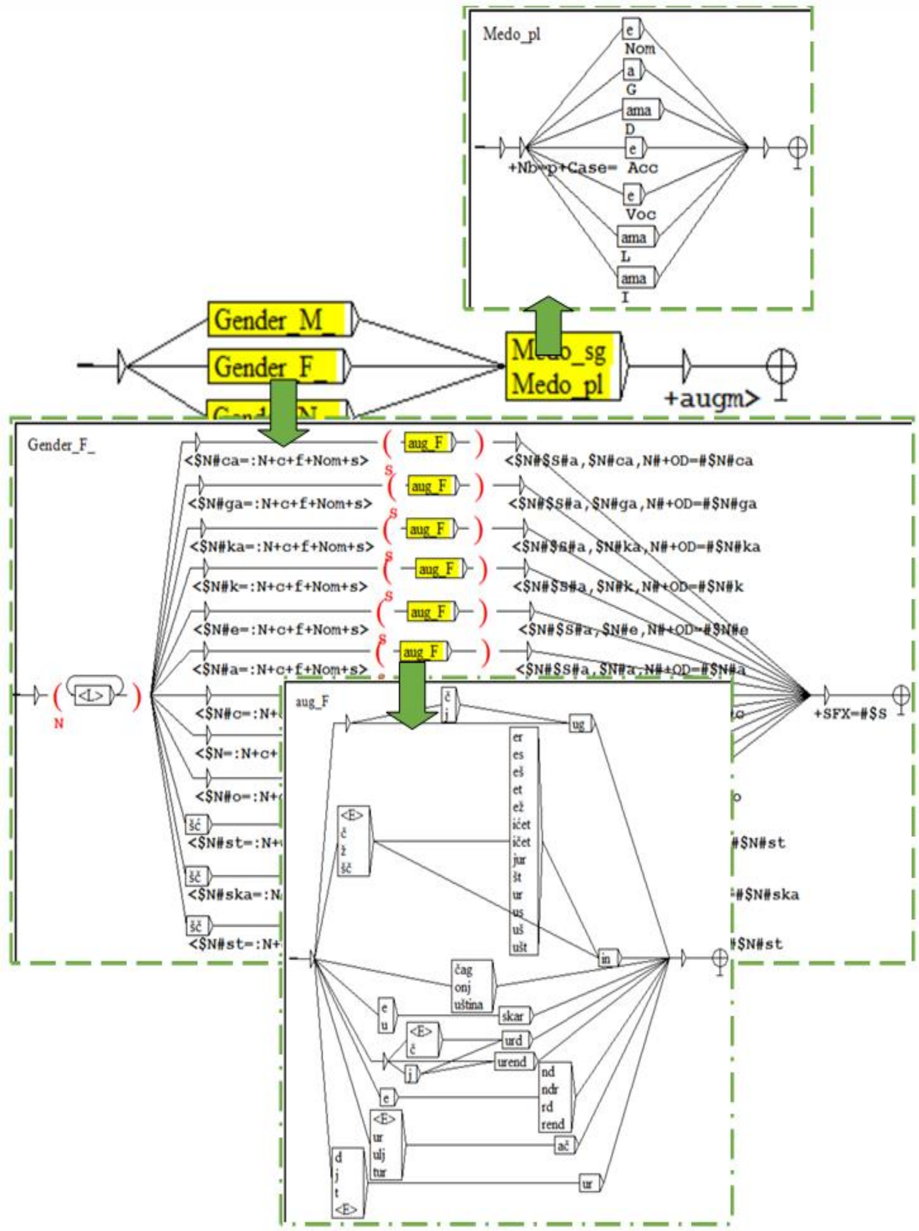

Fig. 2. Section of a morphological grammar recognizing augmentatives

After any of these suffixes is recognized in the current string, the word is annotated as an augmentative noun with the main noun as its superlemma. The case and number information are added in the following step after the string passes through the subgraph Medo_sg or Medo_pl where possible case and number endings are given (Fig. 2.).

Our grammar is not retained from the ambiguity since different nouns may produce the same augmentative (and diminutive) forms. For example, augmentative ***kosturina*** may be recognized from the noun *kost* (en. bone*)* or *kostur* (en. skeleton) and augmentative ***maščurina*** may be recognized from the noun *mast* (en. grease) or *maska* (en. mask). Of course, the same would happen even if we added derivational paradigms directly to our dictionary entries. But, since lexical ambiguity is not something that we can deal with on this (morphological) level of analysis we do not consider this to be a downside of our grammar.

7 Results

We have tested our grammars on a corpus of children stories and novels, and on a corpus that consists of novels that were not written for children. We assumed that the first corpus would have more augmentatives and diminutives, since it is more common to address children in that way. Both corpora were first manually processed, and all augmentatives and diminutives were marked. The results were compared with those found by our NooJ grammars. Firstly, we will present the quantitative results from both corpora. Table 2 presents the results for diminutives, and Table 3 for augmentatives.

Table 2. Analysis of diminutives in both corpora

Name of the work	Category	Tokens	Dim. constructions *(“little house”)*	Diminutives found		% of Diminutives
				Cumulative	NooJ grammar	
Various children’s stories	CAT1	240 616	94	2 195 + 117 lexicalized	549 unique found, 538 correct	0.96%

Name of the work	Category	Tokens	Dim. constructions *("little house")*	Diminutives found		% of Diminutives
				Cumulative	NooJ grammar	
Cumulative for children's novels (CAT1)		**240 616**	**94**	**2 312 (679 unique)**		**0.96%**
J. Polić Kamov: Isušena kaljuža	CAT2	104 585	52	147	255 unique found, 241 correct	0.14%
J. Kozarac: Đuka Begović	CAT2	33 992	3	80		0.23%
V. Novak: Posljednji Stipančići	CAT2	366 299	21	105 + 59 lexicalized		0.044%
Cumulative for general novels (CAT2)		**504 876**	**55**	**696 (283 unique)**		**0.14%**

Table 3: Analysis of augmentatives in both corpora

Name of the work	Category	Augmentative constructions ("*big house*")	Augmentatives found		% of augmentatives
			Cumulative	NooJ grammar	
Various children's stories	CAT1	60	100 + 14 lexicalized	75 unique found, 41 correct	0,047%
Cumulative for children's novels		**60**	**114 (63 unique)**		**0,047%**
J. Polić Kamov: Isušena kaljuža	CAT2	13	52	56 found, 45 correct	0,049%
J. Kozarac: Đuka Begović	CAT2	3	7		0,0205%
V. Novak: Posljednji Stipančići	CAT2	54	15 + 87 lexicalized		0,0278%
Cumulative for general novels		**16**	**59 (51 unique)**		**0,032%**

We have separately counted augmentative and diminutive constructions like *velika kuća* (en. big house) and *mala kuća* (en. little house) which are not morphologically diminutives nor augmentatives, but semantically would fit into these categories. Our NooJ grammars did not detect them, as they are not considered augmentatives nor diminutives, but they are noted in the table (Table 2 and Table 3) for reference.

After comparing human annotations with those yielded by our grammars, we have found the following results. Table 4 represents precision, recall and F-measure for children novels corpus (CAT1), while Table 5 represents the same statistical measures for general novels corpus (CAT2).

Table 4. The results for children novels corpus

	Precision	**Recall**	**F-measure**
Diminutives	0.9800	0.8239	0.8952
Augmentatives	0.5467	0.6508	0.5942
Overall	**0.7633**	**0,7373**	**0.7447**

Table 5. The results for general novels corpus

	Precision	**Recall**	**F-measure**
Diminutives	0.9451	0.8516	0.8959
Augmentatives	0.8036	0.8824	0.8411
Overall	**0.8743**	**0.8670**	**0.8685**

From the data in Tables 4 and 5 we can conclude that our grammar for detecting diminutives [f-measure: 0.90] performs better than the grammar for detecting augmentatives [f-measure: 0.72]. And if we are to consider both grammars to function as one system, its scores would be over 0.80 [P: 0.82; R: 0.80; f-measure: 0.81], which we found satisfactory for this first try to dealing with such derivational forms.

So, what have we learned from our data and how can we implement that knowledge? Although our preliminary results are in general satisfactory, an analysis of an error typology is necessary in order to improve the grammar, and consequently, its results. After a thorough evaluation, we can offer the following typology of errors:

1. personal names: first names (*Slavica, Ančica*) and last names (*Ilačić, Krpina*)
2. Toponyms: *Kučine, Harmica*

3. Possessive adjectives: *Katino* (en. Kate's), *maćehino* (en. stempother's)
4. Word forms: *sokak* (**not** a small juice *(sok*=juice)), vrtuljak (**not** a small garden *(vrt*=garden*)*)
5. Typos: *piace (***wrong:** pi+ac+e*), proizvođačaka (***wrong:** proizvođač+ak+a*)*

Personal names can be, as shown in Introduction, be diminutive in their origin (*Slava – Slavica, Ana – Ančica*).However Croatian speakers often do not treat them as diminutive forms, but as regular noun forms, similar to example (*ploča,* panel – *pločica* – bathroom tile). Next to that, not all personal names are diminutives or augmentatives by formation, particularly last names. The same can be said about toponyms. Possessive adjectives formed from feminine nouns are another category that came back as augmentative forms in the first results, although they are clearly not.

In the fourth type of errors, although both words are correctly segmented (nouns '*sok'* and '*vrt'* exist in Croatian dictionary and all the suffixes and case endings are valid), their diminutive forms are not '*sokak*' and '*vrtuljak*' but rather '*sokić*' and '*vrtuljčić*' respectfully. In addition, words '*sokak*' and '*vrtuljak*' also exist in Croatian language but their meanings are '*street*' and '*merry-go-round'*. The reason why these words were recognized by our grammar is that they were not in our NooJ dictionary.

Errors of the fifth type are similar to the previous type of errors in that the words are segmented correctly (there really are words '*pi*' and '*proizvođač*' in the dictionary, and both 'ac' and 'ak' are regular suffixes for diminutives, and 'e' and 'a' are case endings that may appear after these suffixes). Still, these words do not exist in Croatian language and as such are marked incorrect.

The grammar for recognition of diminutives and augmentatives is marked as a grammar with low priority level, which means that if the word is recognized by the dictionary, the grammar will not be applied to that word. Thus, errors of type 1, 2 and 4 are easily solved by adding the missing names, toponyms and other word forms. Similarly, if a grammar for recognizing possessive adjectives is build and applied to the text prior to the grammar for recognizing diminutives and augmentatives, it would solve the errors of type 3. For the last type of errors, however, we do not offer any solution that would be possible on this level of text analysis.

8 Conclusion

In this paper we have presented the NooJ grammar-based solution for recognition of diminutives and augmentatives in the Croatian language. We started this project with a thorough review of the literature in order to analyze the morphology of Croatian diminutives and augmentatives. In the Croatian language, diminutives and augmentatives are formed from the base word with a number of suffixes, such as -ić and-ica for diminutives or -ina for augmentatives. Based upon this data, our grammars were constructed. Their task was to recognize and mark any diminutive or augmentative form, in case that the base word was already in the dictionary (e.g. *stol*, table ->*stolić*, little table). We have tested the grammars on two stylistically different corpora, the one based on children novels, and the one based on general novels in Croatian language.

Except for finding data that helped us evaluate our grammars, we have also collected data about the frequency of diminutives and augmentatives in the Croatian language. In the case of diminutives, we have found them to be most common in children novels, comprising almost 1% of all words in the text, while they accounted for only 0.14% of tokens in general novels corpus. As for the augmentatives, they are a very rare word form in texts, comprising less than 0.05% of tokens in both corpora.

Our grammars have successfully found a large percentage of diminutives, with average precision over 96%, recall over 83% and F-measure of 89%. In the case of augmentatives, the numbers were somewhat lower, due to the large number of homographic words. In the Croatian language, a word can be e.g. a possessive adjective with the suffix –*ina*, which is the most common augmentative suffix. Due to the large number of false positives, the grammar for recognizing augmentative forms yielded the average precision of 67%, recall of 76.6% and F-measure of 71.7%. Although considerably lower than in the case of diminutives, it should be noted that these results are based on a much smaller number of words, as augmentatives are much rarer in Croatian texts.

It is a matter for discussion if this manner of processing diminutives and augmentatives in Croatian is indeed the best method, as we have to take into account many homographic words which need to be differentiated from true results, the non-standard forms, and those based on words which are not already in the NooJ dictionary. We can propose that the best method for a complete recognition system would be both the construction of NooJ grammars, and manual addition of derivational paradigms directly to the NooJ dictionary. However, as we have seen that the relative frequency of diminutives, and especially augmentatives, is considerably low, this manual method may be too expensive and too slow. At this point, the grammars recognize almost all the diminutives and majority of the augmentatives. Thus, it is our strong belief that this work can be of referential benefit to future researchers of Croatian diminutives and augmentatives.

9 References

1. Babić, S.: Sustav u tvorbi hrvatskih umanjenica, (*System for building Croatian Diminutives*). In *Slavistična revija,* Letnik 20:1, Ljubljana (1972) (in Croatian)

2. Babić, S.: Sufiksalna tvorba uvećanica u hrvatskome književnome jeziku, (*Suffix Word Formation of Augmentatives in Croatian Literary Language*). In Suvremena lingvistika*, 41-42(1-2),* pp. 11-20*,* http://hrcak.srce.hr/23938 (1996) (in Croatian)

3. Babić, S.: Tvorba riječi u hrvatskom književnom jeziku. (*Word Formation in Croatian Literary Language*). Zagreb: Hrvatska akademija znanosti i umjetnosti: Nakladni zavod Globus, pp. 215-221 (2002) (in Croatian)

4. Bosanac, S., Lukin, D., Mikolić. P.: A Cognitive Approach to the Study of Diminutives: The Semantic Background of Croatian Diminutives*, Rector's Award* paper, Zagreb: Filozofski fakultet, academic year 2008/2009 (2009)

5. Grandi. N.: Development and Spread of Augmentative Suffixes in the Mediterranean Area. In P. Ramat and T. Stolz (eds.) Mediterranean languages, Bochum, Dr. Brockmeyer University Press, pp. 171-190 (2002)

6. Grandi, N.: Renewal and Innovation in the Emergence of Indo-European Evaluative Morphology. In Körtvélyessy , L., Stekauer, P. (eds.): Diminutives and Augmentatives in the Languages of the World. Lexis: e-journal in English lexicology, 2011, vol. 6, pp. 5-25,http://lexis.revues.orgimg/pdf/Lexis_6.pdf (2011)

7. Hornoiu, D.: Avoiding disagreement in Romanian conversational discourse: the use of diminutives. In A. Bucuresti Editura Univerităţii Bucuresti, pp 99-105(2008)

8. Jovanović. V.: Deminutivne i augmentativne imenice u srpskom jeziku. Beograd: Institut za srpski jezik SANU (2010) (in Serbian)

9. Katunar, D.: Diminutives in Action: A cognitive account of diminutive verbs and their suffixes in Croatian. In Suvremena lingvistika*,* 39(75), pp. 1-23, http://hrcak.srce.hr/105494 (2013)

10. Körtvélyessy L.: A Cross-Linguistic Research into Phonetic Iconicity. In: Diminutives and Augmentatives in the Languages of the World. Lyon, 2011. pp. 27-40, http://screcherche.univ-lyon3.fr/lexis/IMG/pdf/Lexis_6.pdf (2011)

11. Krstev C. and Vitas D.: Extending the Serbian E-dictionary by using lexical transducers. In Formaliser les langues avec l'ordinateur : De INTEX à NooJ. Edited by Svetla Koeva, Denis Maurel and Max Silberztein. Cahiers de la MSH Ledoux. Presses Universitaires de Franche-Comté., France: 147-170 (2007)

12. Marković, I.: Uvod u jezičnu morfologiju. Zagreb: Disput (2013) (in Croatian)

13. Mitrićević-Štěpánek, K.: Deminutivi u funkciji nabrajanja i konfrontacije u češkom i srpskom jeziku, In Opera Slavica XVII, 2007, 4, https://digilib.phil.muni.cz/bitstream/handle/11222.digilib/117151/2_OperaSlavica_17-2007-4_4.pdf?sequence=1 (2007) (in Serbian)

14. Mota C.: Portuguese morphology with INTEX 4.33. In Formaliser les langues avec l'ordinateur : De INTEX à NooJ. Edited by Svetla Koeva, Denis Maurel and Max Silberztein. Cahiers de la MSH Ledoux. Presses Universitaires de Franche-Comté., France: 135-146 (2007.)

15. Panocová, R.: Evaluative suffixes in Slavic. In Bulletin of the Transilvania University of Brasov, Series IV: Philology and Cultural Studies, Vol. 4 (53) No. 1, pp. 175-182 (2011)

16. Pintarić, N.: Kontrastivno rječotvorje: imenička tvorba u tablicama, Zagreb: Filozofski fakultet, Odsjek za zapadnoslavenske jezike i književnosti (2010) (in Croatian)

17. Sicherl, E., Žele, A.: Nominal diminutives in Slovene and English. In *Linguistica (Ljubljana)*, številka letn. 51, pp. 135-142 (2011)

18. Silberztein, M., 2003. NooJ Manual, www.nooj4nlp.net.

19. Spasovski, L.: Morphology and Pragmatics of the Diminutive: Evidence from Macedonian. Doctoral Thesis, Arizona State University (2012)

20. Vučković, K., Tadić, M., Bekavac, B.: Croatian Language Resources for NooJ. In CIT. Journal of computing and information technology, 295-301, 18 (2010)