

# An Inheritance Based Description of Bulgarian Noun Inflection

Katina Bontcheva and James Kilbury

Computerlinguistik

Institut für Sprache und Information

Heinrich-Heine-Universität Düsseldorf

Universitätsstr. 1

D-40225 Düsseldorf, Germany

{bontcheva, kilbury}@ling.uni-duesseldorf.de

## Abstract

The paper discusses a non-monotonic inheritance analysis of Bulgarian noun inflection that uses abstract morphophonemic representation. The analysis is encoded in DATR.

## 1 Introduction

In this analysis, we approach Bulgarian noun inflection in the terms of network morphology<sup>1</sup> and several other closely related inflectional morphology theories. Thus, essential feature of the analysis is that inflectional types are represented as objects with non-monotonic inheritance hierarchy that gives explicit account of regularities, sub-regularities and exceptions through highly constrained lexical entries, default inflectional class(es) and non-productive class(es).

Our goal has been to make a precise and explicit description allowing economical lexical entries. We decided to use abstract deep morphophonemic representation to simplify the description and the morphotactics. For example, the phoneme *i* has at least 3 sources: *i*, *i*<sub>2</sub> and *ι*. The *i* that is not involved in morphonological alternations is transliterated as *i* while the *i* that triggers 2<sup>nd</sup> palatalization is marked in the programme as *2* but surfaces as *i* in the generated forms.<sup>2</sup> This may look like diachronic description but it is not; rather our description is synchronic and application oriented since the use of deep morphophonemic representation is always motivated by alternations at the synchronic level.

<sup>1</sup> Cf. (Corbett & Fraser 1993).

<sup>2</sup> Our abstract Latin transcription for Bulgarian differs from the ISO standard in several ways:

- it makes use of numerals in abstract morphophonemic representations, e.g. *e* < *ĕ*<sub>1</sub> (in endings) is *4*, *e* < *ĕ*<sub>2</sub> (in endings) is *5*, *e* < *ь* (in suffixes) is *6*, *ь* (in suffixes) is *8* etc.

- *ž*, *č*, *š* are represented with the digraphs *zh*, *ch*, *sh* due to restrictions on the diacritics imposed by the formalism DATR

- for the same reason *ѣ* and *ѥ* are transcribed with *y* and *j* (instead of “ and ‘)

- we refrain from using digraphs for the transcription of *ю* and *я* and use *w* and *q* instead because we want to keep as close as possible to Bulgarian orthography.

## 2 Bulgarian nouns

Nouns in Bulgarian have two inflectional categories – number and definiteness, and one structural category – gender.

The most important elements through which the paradigmatic oppositions are achieved are base, sg stem suffixes (suff sg), plural stem suffixes (suff pl), inflections (flex pl) and articles.<sup>3</sup>

### 2.1 Bases

The base of a noun is the part that is left after the inflectional elements such as singular and plural stem suffixes, endings and articles have been removed. Thus, the base in Bulgarian consists of optional prefixes, root(s), and optional suffixes. For example, the word *uchii-shte* ‘school’ splits into the base *uchilisht* and the sg stem suffix *e*. The suffix *isht* belongs to the base as it appears in all forms of the lexeme. The same suffix however, may be treated as an inflectional element, e.g. a plural stem suffix, cf. *kup* ‘heap’ – *kup-isht-a* ‘heaps’ where *a* is the plural ending and *isht* is a plural stem suffix.

### 2.2 Stem suffixes

The sg stem suffixes define the gender of the noun. Every noun in Bulgarian belongs to one of the 3 genders - masculine (masc), feminine (femn) or neuter (neut). Sg stem suffixes are not proper inflections – they are morphomic affixes<sup>4</sup> and are added to the base to produce the singular stem. Since there are no endings for the singular number, the singular stem is equivalent to the form for the singular non-definite.

The gender and some lexical semantic properties (e.g. person (pers) vs. non-person or proper (prop) vs. common noun) of the noun determine which forms are relevant for each lexeme. Thus, feminine and neuter nouns have only one definite form in the singular, while masculine nouns have two definite forms in the singular – one form with the short definite article, and another with the full definite

<sup>3</sup> For details cf. (Gramatika 1983; Gramatika 1991).

<sup>4</sup> For details on morphomic stems cf. (Blevins 2003).

article.<sup>5</sup> Proper names lack definite forms in singular; masculine nouns for non-persons have a distinct form for plural that is used after numerals.

The default sg stem suffix is the zero suffix. In this case the form for the singular non-definite is equivalent to the base. The vast majority of nouns with zero suffix are of masculine gender. They belong to the default inflectional class<sup>6</sup> N (nouns) in our analysis, e.g. *narod* ‘people’, *boec* ‘warrior’ etc. A group of feminine nouns also has the zero sg stem suffix (nouns with the derivational suffixes *-ost/-est*, e.g. *radost* ‘joy’, and a small number of other nouns, e.g. *pesen* ‘song’). They also belong to the default inflectional class N. Another group of nouns that take the zero suffix are the loan words that end in the vowels *i, u, w*.<sup>7</sup> We call the base of such nouns *vlwb* (vocalic loan word base). The majority of these nouns are neuter, e.g. *taksi* ‘taxi’, *menw* ‘menu’. These nouns are also in the default inflectional class N. The few exceptions that denote male persons, e.g. *atashe* ‘attaché’, *parvenw* ‘parvenu’ are masculine but add the inflectional elements for neuter. They build a separate unproductive class together with other masculine nouns that have plural inflection *a* (NM\_a).

The sg stem suffixes *a* and *q*<sup>8</sup> are associated by default with the feminine gender. These nouns are also assigned to the default inflectional class N, e.g. *zhena* ‘woman’, *banq* ‘bathroom’, *Mariq* ‘Maria’. A small group of nouns with sg stem suffixes *a* and *q* denote male persons, e.g. *bashta* ‘father’, *sydiq* ‘judge’. They are masculine but add the inflectional elements for feminine and constitute a separate, unproductive class NMa (nouns, masc, sg suffix *a*).

The sg stem suffixes *o* and *e* are associated by default with the neuter gender. These nouns also belong to the default inflectional class N, e.g. *selo* ‘village’, *pole* ‘field’. A small group of neuter nouns with the sg stem suffix *o* constitute a separate, unproductive class NNo\_i, e.g. *zhivotno* ‘animal’. Another small group of nouns with the sg stem suffix *e* denote male persons, e.g. *dqdo* ‘grandfather’, and are masculine but add inflectional elements for neuter and constitute a separate unproductive class NMo.

Finally, the suffixes *in* and *j* are not proper structural suffixes as *in* is used to build singulative forms of masculine nouns for nationality and ethnicity, and *j* formally signals masculine stems that show palatal alternations. The nouns with *in* belong to the default inflectional class N, and the nouns with *j* belong to the same class or to the less productive NM\_e class.

Plural stem suffixes are added to the base to produce the plural stem but do not contribute to the formal classification of the nouns in genders. Pl stem suffixes *et, en, es* are associated with neuter nouns that had old consonant stems. However, *et* became very productive and

spread over most of the disyllabic neuter nouns with sg stem suffix *e*, and its reduced variant *t* is now used to build the plural stems of loan words that end in *i, u, w, e* (classes N and NM\_a). Pl stem suffix *ov* (< \*-ew-) is associated with the masculine nouns that belonged to the old *u*-declension, which also became very productive and spread over most of the monosyllabic masculine nouns. Pl stem suffixes *isht*, and *ovc* (< *-ov-ec* with elision of the suffixal *e* (< ы) before the plural ending) are used to combine in one paradigm the singular form of one lexeme, e.g. *pyt* ‘road’, *chernodreshko* ‘a man with a black folk outfit’ with the plural forms of another (most often non-existent lexeme – *\*pytishte, \*chernodreshkovec*) – *pyt-isht-a, chernodreshk-ovc-i*.

## 2.3 Endings

The singular number is not marked by inflection. The endings for the plural are:

- *i* (< ы) is the default for the feminine nouns that belong to N class (and e.g. for adjectives)
- 2 (*i* < *i*<sub>2</sub> that triggers 2<sup>nd</sup> palatalisation) is the default for masculine nouns that belong to the N class, e.g. *nalog* ‘tax’
- *a* is the default for neuter nouns in the N class, and is used also for the masculine nouns of NM\_a class
- *e* is the inflection for the masculine nouns of NM\_e class, e.g. *most* ‘bridge’, *myzh* ‘man’
- 5 (*e* < *ě*<sub>2</sub> that triggers 2<sup>nd</sup> palatalisation) in *ryce* ‘hands’, *noze* ‘feet’ is the old nom dual ending
- 1 (*i* < *i*<sub>1</sub> that triggers 1<sup>st</sup> palatalisation) is the inflection for the neuter nouns *oko* ‘eye’, *uxo* ‘ear’.

The inflection for the numeric plural is *a*, which is added to the singular stem to build a special form of masculine nouns for non-persons. This form is used after numerals and some pronominal quantifiers such as *nqkolko* ‘some’. The source of this inflection is the ending for nom/acc dual form of the nouns that belonged to the old *o* and *jo* declensions.

## 2.4 Phrasal Affixes

Finally, the definite articles are added to the corresponding non-definite forms. There are good reasons to call the definite articles *phrasal affixes*<sup>9</sup> rather than inflections or endings as they are not bound only to the noun stems but are likely to drift to the leftmost of the NP, attaching themselves to the first nominal element of the NP. Furthermore, they are not proper clitics as they do not always agree in gender with the noun. Rather, there is phonological harmony between the noun and the article,<sup>10</sup> e.g. the masculine nouns that end in *a* and *o* take the articles for feminine, resp. neuter (*sluga-ta, dqdo-to*).

<sup>5</sup> The form with the full article is required when the noun is a subject or a subject complement; otherwise the form with the short article is used.

<sup>6</sup> For more details cf. below.

<sup>7</sup> In our notation *w* transcribes *ю* (ju).

<sup>8</sup> In our notation *q* transcribes *я* (ja).

<sup>9</sup> On clitics and phrasal affixes cf. also (Corbett 2000: 152).

<sup>10</sup> For detailed argumentation against the clitic nature of the article cf. (Ortmann 2003: 8ff.).

There are two articles for masculine singular: *y*,<sup>11</sup> called the *short article*, and *yt*, called the *full article*. As was mentioned above, the form with the full article is required when the noun is a subject or a subject complement, and the form with the short article is used in all other cases. Both forms arise from the nom sg masc of the weak demonstrative pronoun *mъ*. The vowel (*y* in our transliteration) in both articles represents the vocalised final *er* (ѣ) of the singular stem of the noun, and only the *t* in the full article comes from the demonstrative pronoun. The article for feminine singular is *ta* (from the nom sg fem of the weak demonstrative pronoun) and for neuter singular *to* (from the nom sg neut). There are two variants of the definite article for the plural: for neuter nouns *ta* (from nom pl neut of the weak demonstrative pronoun) and *te* (< *tě*<sub>2</sub>) for masculine and feminine nouns instead of the expected *mu* (masc nom pl) and *my* (fem nom pl) since both were influenced by the vowel in the forms for indirect cases.

### 3 Inflectional classes

Now that all important elements have been discussed, we can describe in detail the inflectional classes of Bulgarian nouns as defined in our DATR<sup>12</sup> encoding of the analysis.

The default inflectional class N includes all nouns that have forms predictable from the information listed in the lexical entries and in the class definition. The lexical entries typically contain:

- the node definition
- English translation of the meaning of the noun
- lexical semantic information that is essential for computing the relevant forms.

In order to handle exceptions and irregularities, they may also contain:

- the gender
- the pl stem suffix if different from the default
- the plural inflection if different from the default
- the vocative form, if relevant.

Class definitions are based on the notions of defaults and inheritance. Thus, by default nouns with sg stem suffixes *zero*, *j* and *in* are masculine, nouns with *a*, *q* are feminine and nouns with *o*, *e* or loan nouns with vocalic stems that end in *i*, *u*, *w* are neuter. The default plural inflection for masculine is 2 (*i* < *i*<sub>2</sub>), for feminine *i* (< *ы*) and for neuter *a*. The default for the articles depends on the stem suffixes rather than on the gender of the noun: *y(t)* comes after the default masculine sg suffixes *zero*, *j* and *in*, *ta* comes after the default feminine sg suffixes *a*, *q* and *to* comes after the default neuter sg suffixes *o*, *e* and vocalic loanword stems, *te* is the default article for plural but *ta* comes after plurals in *a*. For all classes, the singular forms are built on the singular stem, and the plural forms

are built on the plural stem. The numeric plural form is an exception and it is built on the singular stem.

Most morphographemic alternations are handled with conventional two-level morphology transducers, which we have encoded in DATR.<sup>13</sup>

Sub-regular exceptions to the defaults described above are grouped in six classes, four for masculine nouns, one for feminine nouns and one for neuter: NM\_e, NMa, NMo, NM\_a, NFcons, NNo\_i. Here is an excerpt from our DATR description that shows these classes:

```
NM_e: % most, chaj, myzh, car, kon
<> == N
<suff pl> == o v
<flex pl> == e .

NMa: % sluga, sydiq (exc. vladika)
<> == N
<gndr> == masc
<pers> == true
<flex pl> == i .

NMo: % dqdo, Ivan
<> == N
<gndr> == masc
<pers> == true
<suff pl> == o v c .

NM_a: % pyt, kup
<> == N
<suff pl> == i sht
<flex pl> == a .

NFcons: % -ost/-est, pesen, misyl
<> == N
<gndr> == femn .

NNo_i: % i zhivotno, oko, uxo
<> == N
<flex pl> == i . % < jery, only oko, uxo
                                with i1
```

Class NM\_e contains masculine monosyllabic nouns and the disyllabic nouns *ogyn*, *vqtyr*, *centyr* ('fire', 'wind', 'centre') that build their plurals with the exceptional *plural inflection e*. The vast majority of them also have *pl stem suffix ov*. Only 5 nouns have a zero pl stem suffix that is marked in the lexical entries: *myzh*, *knqz*, *car*, *kral*, *kon* ('man', 'prince', 'tzar', 'king', 'horse'). The rest of the information is inherited from the default class N.

Classes NMa and NMo contain *masculine* nouns (and proper names) that denote *persons*. The main particularities include also exceptional *sg stem suffixes a* and *o*. Class NMo also has a *pl stem suffix ovc*. The rest of the information is inherited from the default class N.

Class NM\_a contains masculine nouns with exceptional *plural inflection a*. Most of them also have *pl stem suffix isht*. The following particularities are marked in the lexical entries: *krak*, *rog* ('leg', 'horn') have zero pl stem suffix; *atashe*, *parvenw* ('attaché', 'parvenu') have pl stem suffix *t*. The rest of the information is inherited from the default class N.

<sup>11</sup> In our notation *y* transcribes ѣ.

<sup>12</sup> For details cf. (Evans & Gazdar 1996).

<sup>13</sup> For details cf. (Sproat 1992; Evans & Gazdar 1996).

Class NFcons contains *feminine* nouns with the exceptional sg stem *zero suffix*. The rest of the information is inherited from N.

Finally, class NNo\_i contains neuter nouns with the exceptional *plural inflection i*. The following particulari-

ties are marked in the lexical entries: *oko, uxo* ('eye', 'ear') have exceptional plural inflection *l* ( $i < i_1$ ). The rest of the information is inherited from N.

Table 1 lists the classes together with the corresponding suffixes, inflections and articles discussed above.

class	suff sg suff sg suff pl	flex sg flex pl num flex pl	clit def clit pl def
N	-, j, in; a, q; o, e -, j -, et, t, en, es	- a $i < i_2, i; a$	y(t); ta; to te; ta
NM_e	-, j -, j -, ov, ev	- a e	y(t) te
NMa	-a, q -	- i	ta te
NMo	-, o ovc	- i	-, to te
NM_a	-, j - -, isht, t	- a a	y(t); to ta
NFcons	- -	- i	ta te
NNo_i	o -	- $i < i_1, i$	to te

Table 1: Inflectional classes

## 4 Conclusion

Our description is essentially complete with respect to lexical entries and classes, but we plan to extend it for stress information as well as other inflection.

Furthermore, we plan to use it as the basis for a multi-lingual analysis of inflection in Bulgarian, Macedonian (and possibly other South-Slavonic languages) as well as Russian.

## References:

(Blevins 2003) James Blevins, Stems and paradigms. *Language* 79.4, 737–767.

(Corbett 2000) Greville G. Corbett, *Number*, Cambridge: Cambridge University Press, 2000.

(Corbett & Fraser 1993) Greville G. Corbett and Norman M. Fraser, Network Morphology: a DATR account of Russian nominal inflection. *Journal of Linguistics* 29: 113-42.

(Evans & Gazdar 1996) Roger Evans and Gerald Gazdar, DATR: a language for lexical knowledge representation. *Computational Linguistics* 22: 167-216.

(Gramatika 1983) *Gramatika na sãvremennija bãlgarski knižoven ezik*, t. II. Morfologija, Sofia: Izdatelstvo na bãlgarskata akademija na naukite, 1983.

(Gramatika 1991) *Gramatika na starobãlgarskija ezik*, Sofia: Izdatelstvo na bãlgarskata akademija na naukite, 1991.

(Ortmann 2003) Albert Ortmann, *Kategorien des Nomens. Schnittstellen und Ökonomie*, Tübingen: Niemeyer (= Linguistische Arbeiten 458), 2003.

(Sproat 1992) Richard Sproat, *Morphology and Computation*, Cambridge, MA: MIT Press, 1992.

## 6 Appendix

Some example theorems

Bo6c:

```
<gndr> = masc
<sem> = warrior
<prop> = false
<pers> = true
<form sg> = b o e c
<form sg def> = ***
<form sg def shrt> = b o e c a
<form sg def full> = b o e c y t
<form pl> = b o j c i
<form pl def> = b o j c i t e
<form pl num> = ***
<form sg voc> = b o e c o.
```

Mom8k:

```
<gndr> = masc
<sem> = young man/bachelor
<prop> = false
<pers> = true
<form sg> = m o m y k
<form sg def> = ***
<form sg def shrt> = m o m y k a
<form sg def full> = m o m y k y t
<form pl> = m o m c i
<form pl def> = m o m c i t e
<form pl num> = ***
<form sg voc> = m o m k o.
```

Ivan:

```
<gndr> = masc
<sem> = ivan
<prop> = true
<pers> = true
<form sg> = i v a n
<form sg def> = ***
<form sg def shrt> = ***
<form sg def full> = ***
<form pl> = i v a n o v c i
<form pl def> = i v a n o v c i t e
<form pl num> = ***
<form sg voc> = i v a n e.
```

Carj:

```
<gndr> = masc
<sem> = tsar
<prop> = false
<pers> = true
<form sg> = c a r
<form sg def> = ***
<form sg def shrt> = c a r q
<form sg def full> = c a r q t
<form pl> = c a r e
<form pl def> = c a r e t e
<form pl num> = ***
<form sg voc> = c a r w.
```

Dqdo:

```
<gndr> = masc
<sem> = grandfather
<prop> = false
<pers> = true
<form sg> = d q d o
<form sg def> = d q d o t o
<form sg def shrt> = ***
<form sg def full> = ***
<form pl> = d q d o v c i
<form pl def> = d q d o v c i t e
<form pl num> = ***
<form sg voc> = ***.
```

Oko:

```
<gndr> = neut
<sem> = eye
<prop> = false
<pers> = false
<form sg> = o k o
<form sg def> = o k o t o
<form sg def shrt> = ***
<form sg def full> = ***
<form pl> = o c h i
<form pl def> = o c h i t e
<form pl num> = ***
<form sg voc> = ***.
```