Some Empirical Aspects of Verb Inflection in Bulgarian and German

James Kilbury
Computerlinguistik
Institut für Sprache und Information
Heinrich-Heine-Universität Düsseldorf
Universitätsstr. 1
D-40225 Düsseldorf, Germany
kilbury@ling.uni-duesseldorf.de

Katina Bontcheva
Computerlinguistik
Institut für Sprache und Information
Heinrich-Heine-Universität Düsseldorf
Universitätsstr. 1
D-40225 Düsseldorf, Germany
bontcheva@ling.uni-duesseldorf.de

Keywords: inflectional morphology, nonmonotonic inheritance, DATR, empirical evidence, language acquisition, Bulgarian verb inflection, German verb inflection, abstract morphophonemic representation

Abstract

The paper discusses analyses of Bulgarian and German verb inflection that are based on nonmonotonic inheritance and encoded in DATR. The analysis of Bulgarian also makes essential use of abstract morphophonemic representations. We argue that such analyses of inflectional morphology must reflect and be supported by empirical data on language acquisition and mental lexical representations reported in psycholinguistic studies.

1 Introduction

Since the mid 1980's representation languages based on networks with multiple and nonmonotonic inheritance have been employed in descriptions of inflectional morphology. Typically, nodes denote generalized paradigms or inflectional classes, while the relations between inflectional classes are captured by the hierarchical relations between nodes. In 1989 DATR (cf. Evans & Gazdar 1996) was introduced as a *formalism* for lexical knowledge representation that avoids certain problems such as conflicts of multiple inheritance. Since then it has become widely adopted in computational linguistics.

Network Morphology (NM) utilizes DATR notation and principles of inheritance networks in a *linguistic theory* of inflectional morphology (cf. Corbett & Fraser 1993, Corbett 2000). NM reflects a growing theoretical linguistic interest in inheritance-based morphology and is close to the framework of *Paradigm Function Morphology* (cf. Stump 2001).

In the analyses presented here we treat Bulgarian and German verb inflection in terms of NM and closely related frameworks. Inflectional types are represented as objects within a nonmonotonic inheritance hierarchy that provides an explicit account of regularities, subregularities, and exceptions with highly constrained lexical entries, default inflectional classes, and nonproductive classes. We try to avoid as much as possible the use of virtual classes that do not reflect any concrete inflectional type, and, for Bulgarian, we employ abstract morphophonemic representations to simplify the description of the alternations and morphotactics (for details on our use of such representations cf. Bontcheva & Kilbury 2003). Our analyses are encoded in DATR.

2 Empirical basis of NM analyses

While monotonic classifications of inflectional types or other objects can be deterministically computed (cf. Petersen 2004), nonmonotonic classifications in principle allow arbitrary structuring and require external specification as to what counts as regular or general. Thus, NM-based accounts of inflectional morphology must specify further formal criteria and/or empirical evidence for particular analyses.

Early accounts of NM and other DATR-oriented work tended to neglect the empirical criteria for particular proposals involving class hierarchies (cf. Corbett & Fraser 1993), but recent studies seek to provide empirical justification through corpus investigation of frequencies (cf. Brown et al. 2004). In particular, psycholinguistic data on first language acquisition has been advanced to support hierarchical analyses

of German noun inflection (cf. Cahill & Gazdar 1999, Clahsen et al. 2002).

In this paper we couple psycholinguistic criteria with formal constraints on hierarchies in NM accounts of inflectional classes:

- In accord with the studies on German mentioned above we require that immediate dominance (ID) relations between nodes for inflectional classes mirror the direction of *overgeneralization* in language acquisition.
- Moreover, we extend the previous principle to the *transitive closure* of the ID relation.
- Finally, we assume that overgeneralization invariably involves analogy to a concrete inflectional type, which means that hierarchies should include *no nodes for virtual classe*s not represented by real lexemes.

The principles can be illustrated with a part of our analysis of German verb inflection. We assume a linear ordering of nodes for the following classes with V as root:

- V for *lachen* 'laugh'
- Vn for *mahlen* 'grind'
- Vaba for *rufen* 'call'
- Vabb for *schreiben* 'write'
- Vabc for *singen* 'sing'.

V is the regular inflection with preterite stem in *–te* (*lachte*) and perfect participle in *–t* (*gelacht*). Vn inherits everything from V except the form of the participle, which is built with *–n* (*gemahlen*).

Vaba introduces a distinct stem vowel in the preterite stem (*rufen – rief – gerufen*), while Vabb has the same distinct vowel in both preterite and the participle (*schreiben – schrieb – geschrieben*). Finally, Vabc introduces three distinct vowels (*singen – sang – gesungen*).

Our hierarchy makes very explicit empirical predictions that should be confirmed or refuted by psycholinguistic data. Overgeneralizations in language acquisition are invariably directed upward in the hierarchy, so that we find **gemahlt* but never **gelachen* in data from children (cf. Clahsen et al 2002).

For the preterite *rief* from *rufen* we expect regular **rufte* (V), but *mahlen* should never have a form like **muhl* in analogy to *fuhr* from *fahren* (Vaba).

For *geschrieben* we expect the forms **geschreiben* (Vn/Vaba) and *geschreibt* (V), but *rufen* would not have a form like **geriefen* (Vabb) in analogy to *geschrieben* from *schreiben*.

Finally, beside *gesungen* the analysis predicts the three forms **gesangen* (Vabb), **gesingen* (Vn/ Vaba), and **gesingt* (V). All the predictions ultimately should be confirmed in order to support the analysis. If only the form **gesingt* is in fact attested, then this would speak against the specific hierarchical structuring proposed here.

Note also that our analysis excludes any paradigm not found in a real inflectional type (e.g., *rufen – rief – *geruft*) and accordingly involves no corresponding nodes for virtual classes. This restriction is fulfilled but not explicitly discussed in some recent descriptions of German noun inflection (cf. Cahill & Gazdar 1999, Kilbury 2001).

Further details of this analysis and the empirical evidence will be presented in another study. We break off the discussion here and go on to a parallel investigation of Bulgarian morphology.

3 Bulgarian verb inflection

In this paper we describe the construction of the synthetic forms that involve only inflection, i.e., the forms for present, aorist, and imperfect tense, the participles (present active, aorist, imperfect, past passive), the verbal noun and adverb, and the simple forms for imperative.

Most Bulgarian verbs have forms for both perfective and imperfective aspect although some verbs are *imperfectiva tantum*, i.e., without a form for perfective aspect. Quite often these forms have different aspectual stems and belong to different conjugations. Also, the perfective and the imperfective stems take different sets of nonfinite inflections. For this reason, although they are forms of one and the same lexeme, the perfective and the imperfective stems appear here as independent entries.

Traditionally, one speaks of three conjugations in Bulgarian. Two of them represent old inflectional types, inherited from Old Church Slavonic. They have numerous classes and even subclasses defined by the

relationship between the present-tense stem and the aorist stem. Actually, these conjugations are virtual – no verb is defined as belonging to the 1st or 2nd conjugation without additional reference to the inflectional class. The third conjugation is new and emerged as the process of secondary imperfectivization with the suffixes *-а-(-я-)/-ва-/-ава-(-ява-)/-ува-* became very productive. In contrast to the 1st and 2nd conjugations, it is not virtual and has no classes.

The verbs belonging to the 1st and 2nd conjugations have two basic stems traditionally called present-tense stem and the aorist (infinitive) stem. The verbs that belong to the 3rd conjugation have only one stem, which has no thematic vowel and is equal to the base (i.e., the aspectual stem).

The present-tense stem is used in the following forms: present tense, imperfect tense, the simple forms for imperative, present active participle, imperfect participle, and the verbal adverb. Actually, it is the imperfect-tense stem that is used to build these forms except those for present tense. The imperative is made from the base. The aorist stem is used for aorist, aorist participle, past passive participle, and the verbal noun.

4 A NM description

In our opinion, the definition of temporal stems is redundant. The important difference is between the aspectual stems. Our model shows that both present-tense stems and aorist stems can be derived from the same abstract representation of the aspectual stem, which we call the 'base' in our DATR description of Bulgarian verb inflection. The morphophonemicsics is handled with conventional two-level transducers (cf. Sproat 1992), encoded in DATR.

Our analysis reveals two alternative hierarchical structures of the inflectional classes of Bulgarian verbs.

The default class Va (i.e., athematic verb) shows the highest degree of regularity and involves no morphophonemic alternations of the aspectual stem. In traditional terms, the two main temporal (present-tense and aorist) stems are identical. The 3rd conjugation is athematic.[1]

There are seven inflectional classes that belong to the 1st (*e*) conjugation and two to the 2nd (*i*) conjugation. Here the class-specific features are stored that constitute subregular exceptions to the defaults described in Va. Most classes have subclasses with regard to stress and the type of base (i.e., vocalic or consonant).

All lower nodes by default inherit the features from higher nodes in the hierarchy. However, the inheritance is nonmonotonic and at each level the information inherited by default can be overridden (cf. the code below).

Bulgarian verbs have two aspectual stems described in different entries (cf. above). Each imperfective or perfective stem[2] adds a set of inflectional elements such as:

- stem vowels for the temporal stems (**sv $tens**, i.e., stem vowels for present, aorist, imperfect tense)
- endings for the tense paradigms (**flex $tens $numb $pers**, e.g., endings for present tense, plural, 1st person or endings for imperfect, singular, 2nd person, etc.)
- endings for the participles (**flex $gdnr**, e.g., endings for feminine gender or for plural)
- suffixes for building the nonfinite forms and the forms of the (synthetic) past tenses (e.g., suffixes for building the past passive participle - **suff papapa**).

The relevant forms can be constructed according to the following formulae:

- **Indicative:** base + sv $tens + suff $tens + flex $tens $numb $pers
- **Imperative:** base + flex impr $numb

[1] Most often it is claimed that the stem vowel is *a*, which is not true. The last element of the aspectual stem (the base) is a derivational or imperfectivizing suffix that ends in *a* (*а-(-я-)/-ва-/-ава-(-ява-)/-ува-*). For more details on the athematic nature of the 3rd conjugation cf. Maslov (1982: 226, §223).

[2] Imperfective is the default value as there are verbs that are imperfectiva tantum but no verbs that are perfectiva tantum.

- **Participles:** base + sv $tens[3] + suff $part + flex $gdnr
- **Verbal noun:** base + sv noun + suff noun
- **Verbal adverb:** base + sv impf + suff adv.

Whether certain nonfinite forms for the individual aspect stems in fact exist is determined using auxiliary nodes for conditional definitions. Thus, if the stem is imperfective and transitive it has the full set of inflectional forms. If the stem is perfective the following forms are irrelevant: present active participle, verbal noun, and verbal adverb. If the verb is intransitive the past passive participle is irrelevant. However, if the verb is intransitive but the aspectual stem is imperfective it has a neuter past passive participle.[4]

Morphophonemic alternations are handled with two-level transducers (cf. above).

The default values of the inflectional elements are:[5]

<flex pres sg p1> == m[6]
<flex pres sg p2> == s h
<flex pres sg p3> ==
<flex pres pl p1> == m e
<flex pres pl p2> == t e
<flex pres pl p3> == t

<flex past[7] sg p1> ==
<flex impf sg p2> == "<flex impf sg p3>"[8]
<flex impf sg p3> == e
<flex aor sg p2> == "<flex aor sg p3>"
<flex aor sg p3> ==
<flex past pl p1> == m e
<flex past pl p2> == t e
<flex past pl p3> == a

<flex impr sg> == j
<flex impr pl> == jte

<flex masc> ==
<flex femn> == a
<flex neut> == o
<flex plur> == i

<suff pracpa> == s h t[9]
<suff paacpa> == l[10]
<suff papapa> == n[11]
<suff advb> == j k i
<suff noun> == n e

<suff past> == x
<sv $tens> ==
<sv papapa> == "<sv aor>"
<sv noun> == "<sv papapa>"

For programming convenience, we introduce Vt to stand for the thematic conjugations. This is a virtual node where features common to the classes traditionally called the 1st and 2nd conjugations are stored. They constitute subregular exceptions to the defaults described in V3.

Verbs that belong to the thematic conjugations reveal the following subregularities common to all subregular classes:

<flex pres sg p1> == ö % ö = ӧ
<flex pres pl p1> == m
<flex pres pl p3> == ö t % öt = ӧt
<flex impr sg> == i
<flex impr pl> == ete
<sv pres> == e
<sv impf> == ä % ä = ӓ
<form pres sg p1> == P: < "<base>" + "<flex pres sg p1>" * >
<form pres pl p3> == P: < "<base>" + "<flex pres pl p3>" * >

The following defaults are overridden: the values of the present-tense endings for sg p1, pl p1 and pl p3, the values of the stem vowels and the inflections for imperative, and the formulae for computing sg p1 and pl p3 for the present tense. The present-tense stem vowel is specified as *e*, which

[3] Sometimes the vowel that appears in the past passive participle is different from sv aor. We call it *sv papapa*.
[4] These forms can be used impersonally in the passive voice, e.g. *V legloto mi e spano*, lit. 'It has been slept in my bed = Somebody has slept in my bed'.
[5] The empty string means that no ending /inflectional element is added.
[6] For transliteration we use the ISO standard, except for ӓ, ӧ ('ä, ö').
[7] Imperfect and aorist are referred to as *past* wherever their personal endings are identical.
[8] The endings for sg p2 are referred to the endings for sg p3 as the later forms are more frequent. For details on inflectional syncretism and frequency cf. Brown et all. (2004).
[9] pracpa = present active participle.
[10] paacpa = past active participles, i.e. aorist and imperfect participles. They have identical suffixes.
[11] papapa = past passive participle.

reflects the fact that seven out of nine subregular inflectional classes have this value.

In addition, each subregular class has specific subregularities. These are the node definitions for the verbs of the types *mija* 'wash' and *molja* 'ask, beg, plead':

```
Ve7:<> == Vt
  <sv aor> ==
  <suff papapa> == t
  <sv noun> == e
  <flex pres sg p1> == j ö
  <flex pres pl p3> == j ö t
  <flex impr $numb> == Va.

Vi1:<> == Vt
  <sv pres> == i
  <sv aor> == i
  <sv papapa> == e.
```

Lexical entries typically specify the inflectional class and an English gloss. In order to handle exceptions and irregularities, entries may also specify the aspect and the transitivity of the verb, if different from the defaults, or other information.

In our alternative description (cf. Figure 1), the default class Va is unchanged. The differences lie in the structuring of thematic classes and subclasses. There are no virtual nodes. Node Ve stores all the features from the virtual node Vt in our traditional description and the features typical for the traditional 2nd class of the 1st (*e*) conjugation. These verbs have features that could serve as defaults for thematic verb inflection: accent on the temporal stem vowel, present-tense stem vowel *e*, aorist stem vowel *a*, suffix for past passive participle *n,* and consonant base. Class Ve1 contains verbs that have different accentual type, and Ve3 verbs with the 'jat' stem vowel in aorist. In class Ve2 most of the defaults inherited from Va and Ve are overridden. Further down appear verbs with vocalic bases or the *-t* suffix for past passive participle. The structuring of the Vi classes follows the same principles: higher in the hierarchy are the default accentual types and temporal stem vowels, the *–n* suffix for past passive participle, and the consonant bases.

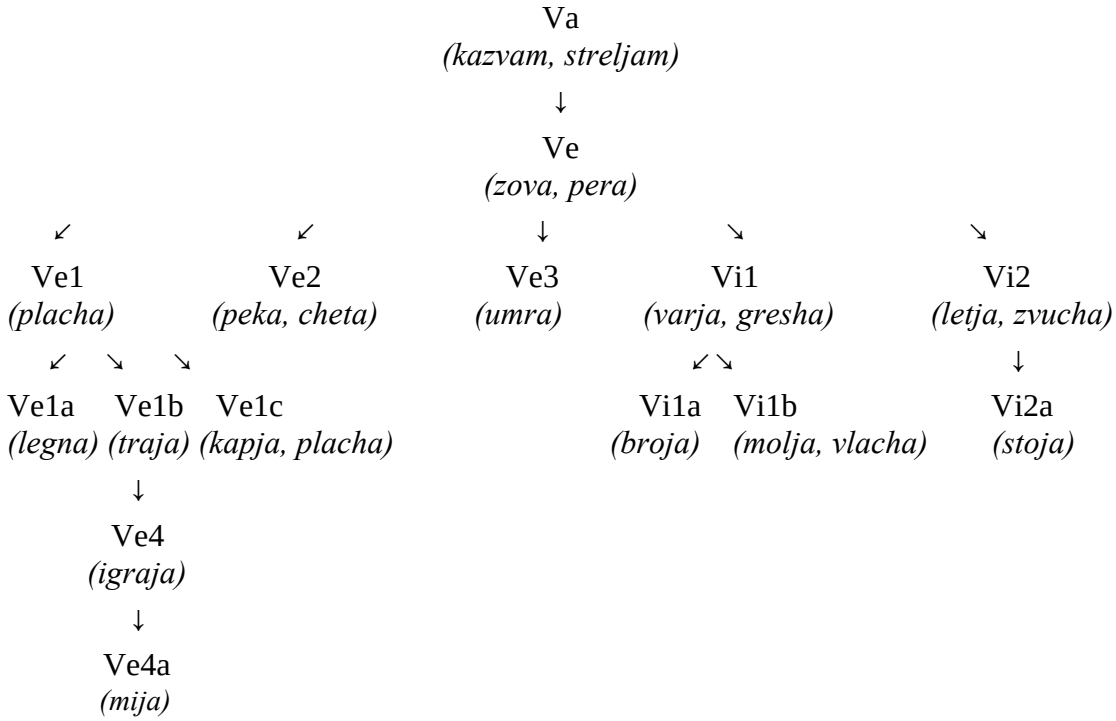

Figure 1: Inflectional classes of Bulgarian verbs (alternative description)

5 Empirical data from psycholinguistics, dialectology, and lexicology

There are several facts that speak in favor of our alternative description.

Loan verbs are assimilated mainly in the 3rd conjugation – our default Va class, e.g. verbs with suffixes *–ira-* and *–(d)isa-: motoriziram, bojadisam*, etc.

Forms like *jax* 'ate', *dax* 'gave' instead of *jadox, dadox* are quite frequent in the speech of children up to the age of 5-6 years. The overgeneralization is from class Ve2 to the root class Va and constitutes a typical example of the athematic aorist.

Most western Bulgarian dialects have only a single set of personal endings for present tense, which is that of Va.

Many verbs of class Ve4a tend to build doublet forms of the past passive participle with *–n*, which is typical for the higher class Ve4: *izpjat* : *izpjan*, *proljat* : *proljan*, *razvjat* : *razvjan*, etc.

6 Conclusion

We have presented and discussed analyses of Bulgarian and German verb inflection that are based on nonmonotonic inheritance and encoded in DATR. Furthermore, we have argued that such analyses must reflect and be supported by empirical data and analyses from psycholinguistic studies. The paper provides evidence for parts of our analyses and highlights further clear empirical claims that must be the object of future investigations.

References

Bontcheva, Katina & Kilbury, James (2003) An inheritance-based description of Bulgarian noun inflection. *Proceedings of the Workshop on Balkan Language Resources and Tools*. Thessaloniki.

Brown, Dunstan, Tiberius, Carol & Corbett, Greville (2004) Inflectional syncretism and corpora. *Proceedings of 5th International workshop on linguistically interpreted corpora (LINC-04)*. Geneva.

Cahill, Lynne & Gazdar, Gerald (1999) German noun inflection. *Journal of Linguistics 35*: 1-42.

Clahsen, Harald, Prüfert, Peter, Eisenbeiss, Sonja & Cholin, Joana (2002) Strong stems in the German mental lexicon: evidence from child language acquisition and adult processing. In Kaufmann, Ingrid & Stiebels, Barbara (eds.) *More than words. A Festschrift for Dieter Wunderlich*. (= Studia grammatical 53), 91-112. Berlin: Akademie Verlag.

Corbett, Greville G. (2000) Formal approaches to Slavic Linguistics. In King, Tracy Holloway & Sekerina, Irina (eds.) *Annual workshop on formal approaches to Slavic linguistics: The Philadelphia meeting 1999* (Michigan Slavic Materials 45), 120-151. Ann Arbor: Michigan Slavic Publications.

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

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

Kilbury, James (2001) German noun inflection revisited. *Journal of Linguistics* 37: 339-353.

Maslov, Yurij S. (1982) *Gramatika na bălgarskija ezik*., Sofia: Nauka i izkustvo.

Petersen, Wiebke (2004) A set-theoretical approach for the induction of inheritance hierarchies. *Electronic Notes in Theoretical Computer Science* 53: 1-13.

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

Stump, Gregory T. (2001) *Inflectional Morphology. A Theory of Pradigm Structure.* Cambridge: Cambridge University Press.