Data-set "New Bulgarian Data" (NBD)

Thomas Zastrow

March 13, 2006

Contents

1 Introduction

In the dialectometry-project, we have two different data-sets: The one which was used by Petya Osenova and a new one, assembled by Prof. Zhobov and his colleagues.
To make a clear distinction between them, the term "New Bulgarian Data" (NBD) is used for the last one.
The NBD-data-set is developed step-by-step. When starting this document, 20 sites were available. The results in the next chapter will change when the data-set grows, so here is the actual date:

13. March 2006, 43 sites available

The file *villages-coord-XML.xml* contains the geographical coordinates of the sites.

2 Some Statistics

The following statistics are compiled by the Java-Class *examination.java*. In the most cases, it sends a X-Query to the eXist-database and does something with the result.
For a better understanding, the X-Qeries are shown here, together with a short explanation.

2.1 Query: Getting all key-tags:

```
collection('/db/nbd')//key
```

Result: There are 6676 keys in NBD. These consists of 161 different keys. (When using the english translation, there are just 159 different tags.)

2.2 Query: Getting all documents in NBD

```
for $child in xmldb:get-child-resources("/db/nbd/")
order by $child
return $child
```

The title of the documents[1] can be used to examine how many entries are in the single files:

[1] The query gives you also back the document with the coordinates, villages-coord-XML.xml. It can be ignored unless you'll need some geographical information.

2.3 Query: How many entries are in the files:

```
doc("/doc/db/nbd/" + $child)//entry
```

Result: Take a look at the appendix

2.4 Query: Getting all variants

```
collection("/db/nbd/")//sampa/variant
```

Result: There are 7491 variants in the data-set. The query gives them in the form:

```
<variant ana="Ncnsi">‘7gni</variant>
```

The attribut "ana" is optional, take a look at the README-file in the appendix. For the further processing, the variant-tags are deleted, together with the attributes. The resulting list is written to the file "variants.dat".

The 3495 variants are consists of 41525 X-Sampa-Codes.

2.5 Query: How many X-Sampa-Codes containing "ZZZZ"

```
collection("/db/nbd")//variant[contains(., "ZZZZ")]
```

Result: 63

A The README-file (by Petya Osenova)

The data-set comes along with a README-file:

```
Each site has its number and name.

The entry consists of the following elements:

- key:  the cyrillic form without the stress

- cform:  the cyrillic form with the stress


NB!:  cform element has an attribute 'ana' with the morphological tag(s)

The coding of the tags is accessible at:  http://www.bultreebank.org/TechRep/BTB-TR03.pdf


- english:  the word translated into English

- nform:  the working transliterated form

- variant:  the variant(s) of the word for this site

The variant also has the 'ana' attribute

- sampa:  it consists of the normalized form and the variants converted in X-SAMPA.

The variant also has the 'ana' attribute.


- Within X-SAMPA the string ZZZZ means that there is no variant in this site for this word

- Within X-SAMPA the string, which begins with the asterics symbol * means that the variant is inferred


NB!:  When some variant has attribute 'gd="1"', this means that the variant is a wordform of the

lexeme that is with a different grammatical information with respect to the cform.

For example, if the word is 'glava' (head), which is nominative, its true variant will

be also in nominative (glav@).  But if there is some another case form registered, then it gets

the attribute 'gd'(=different grammar).  For example, 'glavu'(accusative).  Note that such forms

come ONLY AS ADDITION to the true forms.
```

B The english words

NBD Alphabet

lamb	one's, self
alive	ours
already	outside
apple	oven
apples	ox
ash	pay-1st singular
bake-1st singular	peak
barley	person
black-fem	pick up-3rd plural
black-masc	picking-up;gathering-verbal noun
black-neut	pocket
black-plural	pretty-masc
blood	pretty-neut;good-neut
bottom	quickly
bovine animal, beef	rain
bread	read-1st singular
carry-3rd plural	red-masc
chain dance	river
cheese	road
cherry	rooster
child	salt
dark-neut	sand
day	Saturday
deep	saw-1st singular
does not want-3rd singular	sheep-pl
down	sheep-sg
dry	shepherd
ear	shepherds
earth	sickle
easy	sister
eat-1st singular	sit-1st singular
eat-2nd singular	something
egg	son-in-law;brother-in-law
eggs	star
enter-1st singular	stone
evening	such
eyebrow	sunday
fear	ten
fire	the man
first-he	the middle
flay-1st singular	the milk
flour	the town
fountain	then
Friday	there is no, will not
gave-3rd plural	thin-neut;thinly
give back-1st singular	this-neut
grapes	those
grass	time
guests	today
hand	tomorrow
hands	tongue
harvest	tree
has come-3rd singular	two
he-goat	up
head	wait-3rd plural
healthy	walnut

Page 1

NBD Alphabet

her	was-3rd person
her-accusative, short form	water
her-dative, short form	we
horse	Wednesday
hungry	well-adv
I	were-1st plural
in	where
inside	which-neut
iron	white-plural
key	whole
lentils	will, shall
lived-3rd plural	wind
man	wine
meat	with
Monday	wolf
month	woman
mother	wool
much;many	yard
name	yellow
night	yesterday
now	you-plural
old man	yours
one-masc	yours-plural
one-neut	

Page 2

C The entries

File	Entries
ustovo-v2.txt.xml	155
vinarovo-v2.txt.xml	155
babjak.txt.xml	155
belica.txt.xml	155
kreta.txt.xml	155
aldomirovci-v2.txt.xml	156
beglezh-v2.txt.xml	154
bogdanov_dol-v2.txt.xml	154
golemo_malovo-v2.txt.xml	155
golica-v2.txt.xml	156
gradec-v2.txt.xml	155
huhla.txt.xml	155
petarnica-v2.txt.xml	156
sekirovo-v2.txt.xml	154
stakevci-v2.txt.xml	157
dolna_studena.txt.xml	155
devenci.txt.xml	155
drabishna.txt.xml	154
zanozhene.txt.xml	155
bansko.txt.xml	155
dragodanovo.txt.xml	155
chernogorovo.txt.xml	155
ezerovo.txt.xml	161
kozichino.txt.xml	155
zamfirovo.txt.xml	155
govedarci.txt.xml	155
merichleri.txt.xml	155
momchivovci.txt.xml	155
momkovo.txt.xml	155
nova_nadezhda.txt.xml	155
opan.txt.xml	155
shiroki_dol.txt.xml	155
smochevo.txt.xml	155
borisovo.txt.xml	155
starmen.txt.xml	155
gabare.txt.xml	155
vranilovci.txt.xml	155
sredec__zlgr.txt.xml	155
zdravkovec.txt.xml	155
dobarsko.txt.xml	155
varbovo-v2.txt.xml	159
kalipetrovo-v2.txt.xml	155
tihomirovo.txt.xml	155