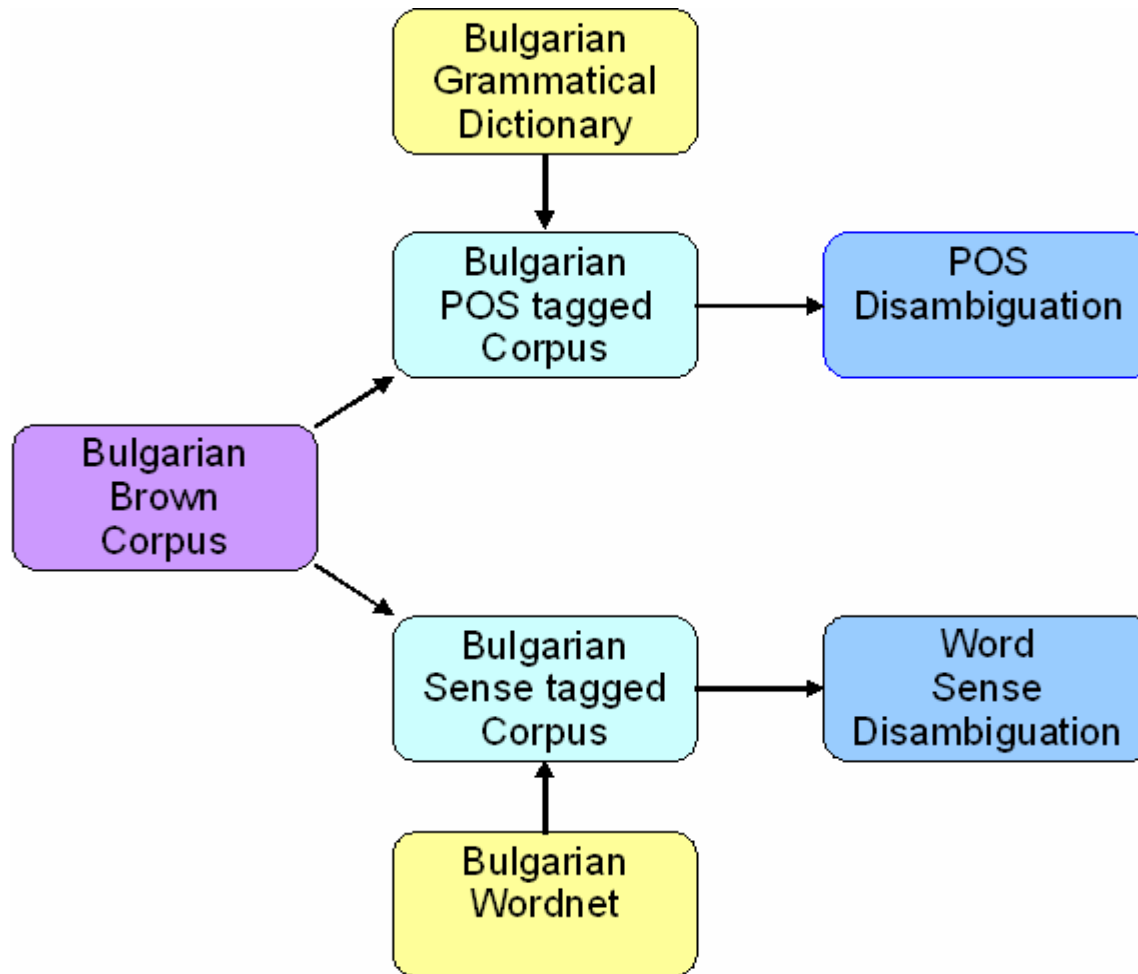


Inflection Morphology of Bulgarian MWEs

Svetla Koeva

Department of Computational Linguistics
IBL – Bulgarian Academy of Sciences

The research framework



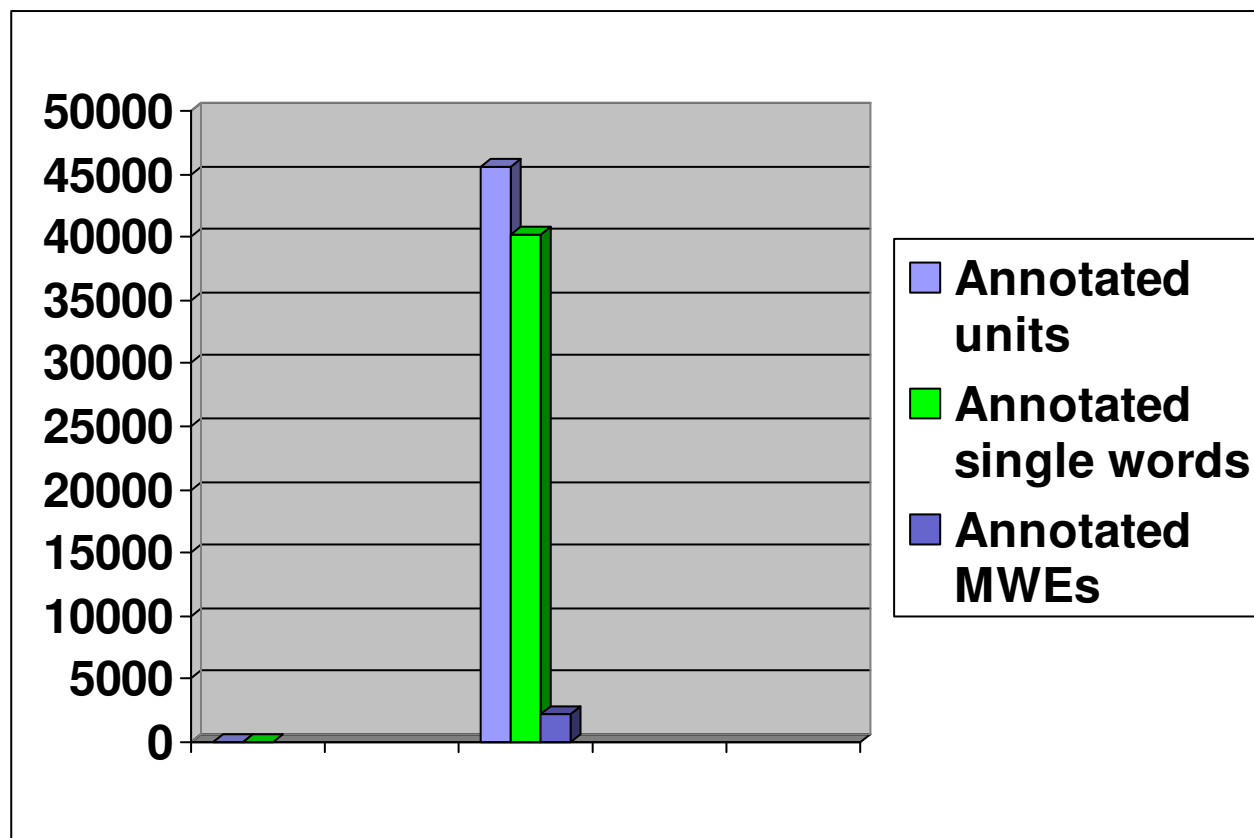
“Brown” Corpus of Bulgarian

- 500 corpus units of approximately 2000 words each, distributed proportionally to language use in 15 categories, containing 1 001 286 words.

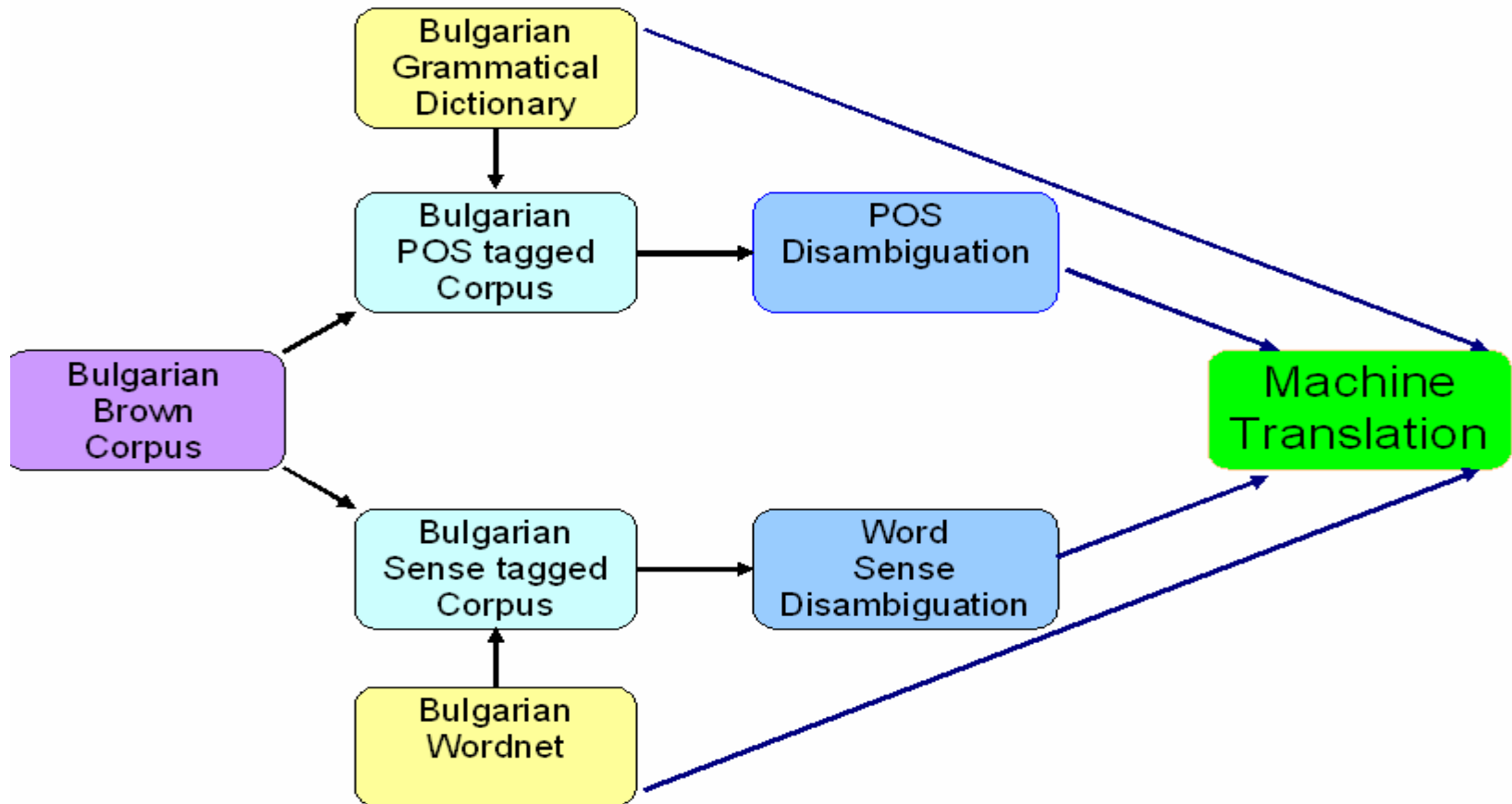
BulNet

- 27 045 synsets, containing 57 496 literals, average number of literals per synset is 2.12.
- Seventeen language-internal relations with 48 371 occurrences, the average number of relations per synset is 1.79.
 - Structure preserving principle
 - Language specific coverage

Sense tagged corpus



The research framework



Resources and Tools in focus

- MWEs inflectional dictionary
- Extension of the amount of the English-Bulgarian parallel corpora
- Parallel corpora for more languages
- Improvement of the Syntactic parser of Bulgarian
- Extension and Cascade model of Morphosyntactic and Syntactic transfer rules
- ...

Quantity of MWEs

- 14 066 compound literals out of 57 496 in BulNet (24,49%);
- 45 769 compound literals out of 203 147 existing (22,5 %) in Princeton wordnet;
- The distribution of multiword expressions (MWEs) among natural languages is approximately equivalent and covers one fourth of the lexis.
- 2 200 out of 46 000 annotated words are MWEs (5,1% quantity; 10,8% simple words coverage).

Definition of a MWE

- MWE – a sequence of two or more graphical words that denote an unique and constant **concept**: Milky way; vitamin A1
- Word – a graphical word that denotes an unique and constant **concept**: chair; vitamin
- Graphical word – a token containing at least one letter or digit: a; chair; milky; way; vitamin; A1
- Token – a sequence of characters between blanks: ...; a; chair; milky; way; vitamin; A1

What is a concept?

- Concept → The general definition of this notion will say that the concept is what we grasp when we understand an expression.
- Concepts are language independent, not observable, and not measurable.

Identifying concepts

- Word – a graphical word that denotes to an unique and constant concept
- A MWE denotes a concept iff a relation of equivalence exists between it and a word from the same language.
{Sofia:1, Bulgarian capital} ‘capital and largest city of Bulgaria located in western Bulgaria’
- A MWE denotes a concept iff a relation of equivalence exists between it and a word from a natural language.
{church officer:1} ‘a church official’
{църковнослужител:1, църковен чиновник:1}

Semantic Ambiguity

- 91% of the MWEs in Bulnet are unambiguous

{communication system:2.

communication equipment:1}

‘facility consisting of the physical plants and equipment for disseminating information’

{communication system:1} ‘a system for communicating’

Inflection dictionaries

- A list of lemmas with marked accent where each entry is associated with a label.
- The label itself represents the grammatical class and subclass to which the respective lemma belongs and contains a number that shows the grammatical type.
- Each label is connected with the corresponding formal description of endings, sound and accent alternations.

Information structure design

- Category information –
6 classes: Noun, Verb, Adjective, Pronoun, Numeral, Others (Adverb, Preposition, Conjunction, Particle, Interjection);
- Paradigmatic information –
Personal, Transitive, Perfective, Common, ...;
- Grammatical information – Inflection, Conjugation, Sound alternations,

Category information

BGD

Characterizes the lemma and shows the grouping of words into grammatical classes.

BDMWEs

Describes the lemma(s) of the head word(s) and indicates the compounds' clustering into grammatical classes.

Noun

- Singular non definite
- Singular definite short article
- Singular definite long article
- Singular definite
- Vocative
- Plural non definite
- Plural definite
- Counting

Paradigmatic information

BGD ● Characterizes lemmas and shows the grouping of words into grammatical subclasses.

BDMWEs ● Characterizes the lemma of the head word(s) and shows the grouping of MWEs into grammatical subclasses;
● Determines the number of other constituents and their word classes;
● Determines the type of blanks.

Lemma of the head word

N+F+I

- Singular non definite
- Singular definite short article
- Singular definite long article
- Singular definite
- Vocative
- Plural non definite
- Plural definite
- Counting

Noun+Feminine+Inanimate

boksova rakavitsa

boksovata rakavitsa

boksovi rakavitsi

boksovite rakavitsi

boxing gloves

Classes of other constituents

- boksova rakavitsa boxing gloves aN
- sazvezdie Rak cancer nN
- hartiya za pisma writing-paper Npn
- sastezavam se compete Vpart
- po sluchay on the occasion of Pn
- ...

Blank type 1

- Always empty

{visshi bozaynitsi:1, placentni bozaynitsi:1}

{placental:1, placental mammal:1, eutherian:1, eutherian mammal:1}

‘mammals having a placenta; all mammals except monotremes and marsupials’

Blank type 2

- Only clitics can appear

{sastezavam se:1; sarevnovavam se:1}

{compete:1, contend:1}

‘compete for something; engage in a contest; measure oneself against others’

sastezavam se

sastezavam **li** se

Blank type 3

- Particular word classes and / or syntactic classes can appear

{snimam:1, **pravya snimka:1**}

{photograph:1, snap:12, shoot:9}

‘record on photographic film’

Pravya **hubavi** snimki.

Shte pravya **tazi sedmica** snimki.

Grammatical information

BGD

Characterizes the formation of word forms and shows the grouping of words into grammatical types.

BDMWEs

Determines the formation of word forms of each component and shows the classification of words into grammatical types.

Determines the fixed and paradigmatically fixed word order.

Determines the agreement dependencies.

Formation of word forms

- Singular non definite лѸмфна жлез^
- Singular definite short article лѸмфна~~та~~ жлез^
- Singular definite long article лѸмфна~~та~~ жлез^
- Singular definite лѸмфна~~та~~ жлез^
- Vocative лѸмфна~~та~~ жлез^
- Plural non definite лѸмфни жлезѸ
- Plural definite лѸмфните жлезѸ
- Counting лѸмфни жлезѸ

Fixed word order

- Fixed word order
 - Combines with blanks type 1 and 2
rod Equus genus Equus
- Paradigmatically fixed word order
 - Combines with blank type 2
poshtenska kutiya post box a(posspro)N
poshtenskata **mi** kutiya
poshtenskata **ti** kutiya
poshtenskata **mu** kutiya

Free word order

- (relatively) Free word order

- Combines with blank type 3

padam duhom

despond

Duhom ne padam nikoga.

Agreement dependencies

- No agreement
 - Combines with blanks type 1, 2 and 3
 - Combines with fixed, paradigmatically fixed and free word order
krasltvo Tonga kingdom of Tonga
- Agreement with the head word
 - Combines with blanks type 1 and 2
 - Combines with fixed, and paradigmatically fixed word order
shopska salata shopska salad
- Independent agreement
 - Combines with blank type 3
 - Combines with free word order
pravya zhest make a gesture
pravya zhestove

BDMWEs – Dictionary format

в'исши боз'айници, a1N+M, 21

з`ахарна ф`абрика, a2N+F, 11

подд'ържам дист'анция, V+N+T3np, 2

с'ъстезавам се, V+N+I2rp, 33

х'окей на лед, N1p1p, 1

BDMWEs – Inflective types format

- Flextype<#>a+2-1-N+F+1+2
- s0<#>" "
- sd<#>'та' "
- v<#>
- p0<#>'1и' '1и'
- pd<#>'1ите' '1и'

Example

	aNF,11	NF,11
Singular non definite	zaharna	fabrika
Singular definite short article		
Singular definite long article		
Singular definite	zaharnata	fabrika
Vocative		
Plural non definite	zaharni	fabriki
Plural definite	zaharnite	fabriki
Counting		

Conclusions

- MWEs represent a large part of vocabulary thus for the purposes of NLP a complete formalization of the inflection of MWEs literals is necessary.
- Comparing to lemmas MWEs have their own inflective rules. Work still remains to be done for the inflectional description of MWEs.