PARSEME

[LEX] **Lexicon issues**

27 - 30. 06. 2016, La Rochelle, France

- Supervisors: Adam Przepiórkowski & Agnieszka Patejuk
- Trainees: Eduard Bejček, Made Windu Kesiman, Lauma Pretkalniņa, Dage Särg, Maria Todorova

Discussion issues:

§ Lexicon model for MWEs to support transcription of ancient manuscripts (Kesiman)

§ Encoding the variability of MWEs (Bejček, Todorova, and Pretkalniņa)

§ Productivity of different elements in a MWE (Särg)

Lexicon model for MWEs to support transliteration of ancient manuscripts (Kesiman)

§ **Problem stated:**

- Complexity of the script (many-to-many relation between compound character class and syllable)
- Writing style (no space between words)
- Limited lexicon (from limited sample of manuscript collection)

§ **Offered solutions:** Using N-gram script characteristics to detect the OCR error

http://www.impact-project.eu/home/

http://www.digitisation.eu/

Publications of the IMPACT leader of the Bulgarian team - Stoyan Mihov

http://lml.bas.bg/~stoyan/lmd/Publications.html

Encoding the variability of MWEs

(Bejček, Todorova and Pretkalniņa)

§ **Problem stated:** a good MWE lexicon should allow to encode:

- What variations on word order are allowed?
  - From fixed (easy) to free (feasible) or almost free (hard to encode)
- What kind of slots an MWE has?
  - Slots as a part of MWE structure can be filled with words from certain syntactic (NP, PP, ...), semantic (first, second, third…) or synonym class/es
  - Slots which allow insertion of elements from context
- What kind of paradigmatic restrictions MWE has?
  - If this is a verbal MWE, does any verb form can be used?

Encoding the variability of MWEs

§ Types of MWE variability

- Paradigmatic - changes in the forms of MWE components
- Syntagmatic - all quantitative and positional changes of the idiom components
- Lexical - constraints on words that could fill a specific slot in a MWE

§ Offered solution - descriptions based on finite-state automata

The framework can be provided with programmes like Unitex, Nooj or Intex

Descriptions based on statistical approaches or semantic clustering.

Encoding the slot variations in MWEs

§ Types of slot variations:

- a synonym *(UN/United Nations Security Council)*
- a close variant *(supporting actor/actress)*
- a (?semantic) group *(CZ: stát v popředí/na výsluní* (to stand in front/in the sun) "to be important"*)*
- an infinite group *(... first/second/seventy-fifth/middle/last/...)*

§ Offered solution - description based on finite-state automata (i.e. regular expressions)

Still: problem with specification of meaning of a new MWE

Encoding the paradigmatic variability of MWEs

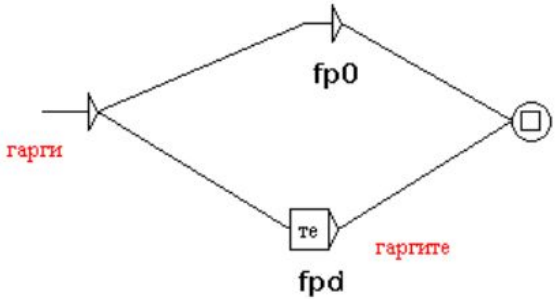
§ The "defective" components' forms (some form/s cause loss of the idiomatic interpretation). For example in Bulgarian for: *pisano e* (it's written) 'it is predetermined' - only the passive form has idiomatic interpretation - compare with paradigmatic forms of the free verb

pisha, pishesh, pishe, pishem, pishete, pishat - Present t.
pisano - pass p.-neut; pisana - pass-fem.; pisan - pass-m.;
pisani - pass-pl.

§ Offered solution - descriptions based on finite-state automata

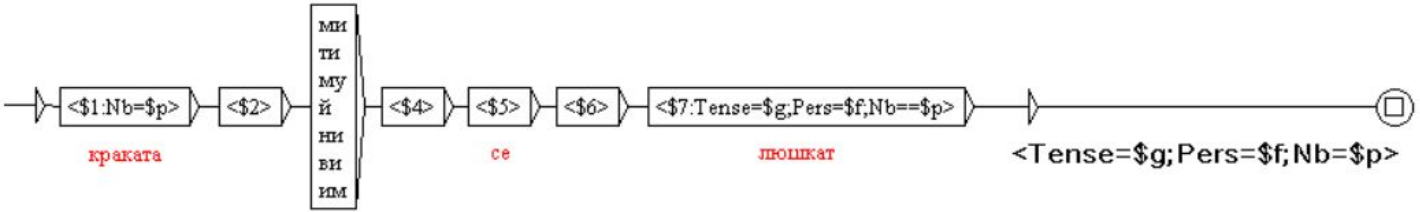
Encoding the paradigmatic variability of MWEs

§ Some approaches with automata



- Inflective descriptions with “defective paradigms”, illustrated for *gargite* in *broya gargite* (count crows) ‘distract’

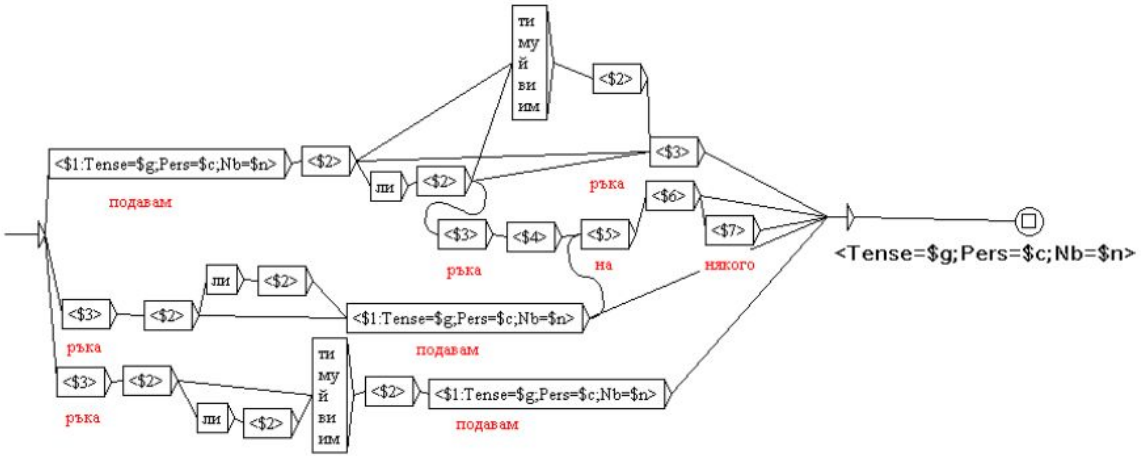
- Descriptions with restrictions inside the syntactic grammar graph



- Special markers inside the free verb paradigm

Encoding the word order variability of MWEs

§ Offered solution - descriptions based on finite state automata, classification of MWE separators (Koeva 2006; Savary 2005)



Problem: encoding almost free word order gives large, complex automaton

Combination productivity of different elements in a MWE (Särg)

§ **Problem stated:** how to compare the productive combinations of different adverbs in adverb-adjective sequences, taking into account the frequencies of adverbs?

Currently:

$$\text{Prod(adv)} = \frac{\text{Number of unique phrases}}{\text{Number of all phrases}}$$

Prod('very') = 3/35 ~ 0.09

Prod('quite') = 2/6 ~ 0.33

Prod('really') = 1/1 = 1

=> the most frequent adverbs come out as least productive

Very good	20
Very bad	10
Very nice	5
Quite bad	3
Quite good	3
Really nice	1

Combination productivity of different elements in a MWE

§ Offered solutions

- Look only the absolute numbers, not taking frequencies into account: Prod('very') = 3, Prod('quite') = 2, Prod('Really') = 1 ... *against the intuition*
- Use Wordnet for finding the number of different synsets and to group them according to number of senses (in case of Estonian adjectives and adverbs, not available)
- Divide by a flattened number of all occurrences
- Ignore adverbs with frequencies lower than a threshold
- Use something like inversed TF-IDF to look at adjectives and adverbs in texts, use flattened frequencies

Very good	20
Very bad	10
Very nice	5
Quite bad	3
Quite good	3
Really nice	1

CONCLUSION

Such a nice group, do not dare to criticize, as they resolve everything with

Thank you!

Eduard Bejček bejcek@ufal.mff.cuni.cz

Made Windu Kesiman madewinduantara.kesiman@gmail.com

Lauma Pretkalniņa lauma.pretkalnina@gmail.com

Dage Särg dage_009@hotmail.com

Maria Todorova maria.todorova@gmail.com