Corpus Management System: Search Functionality Implementation

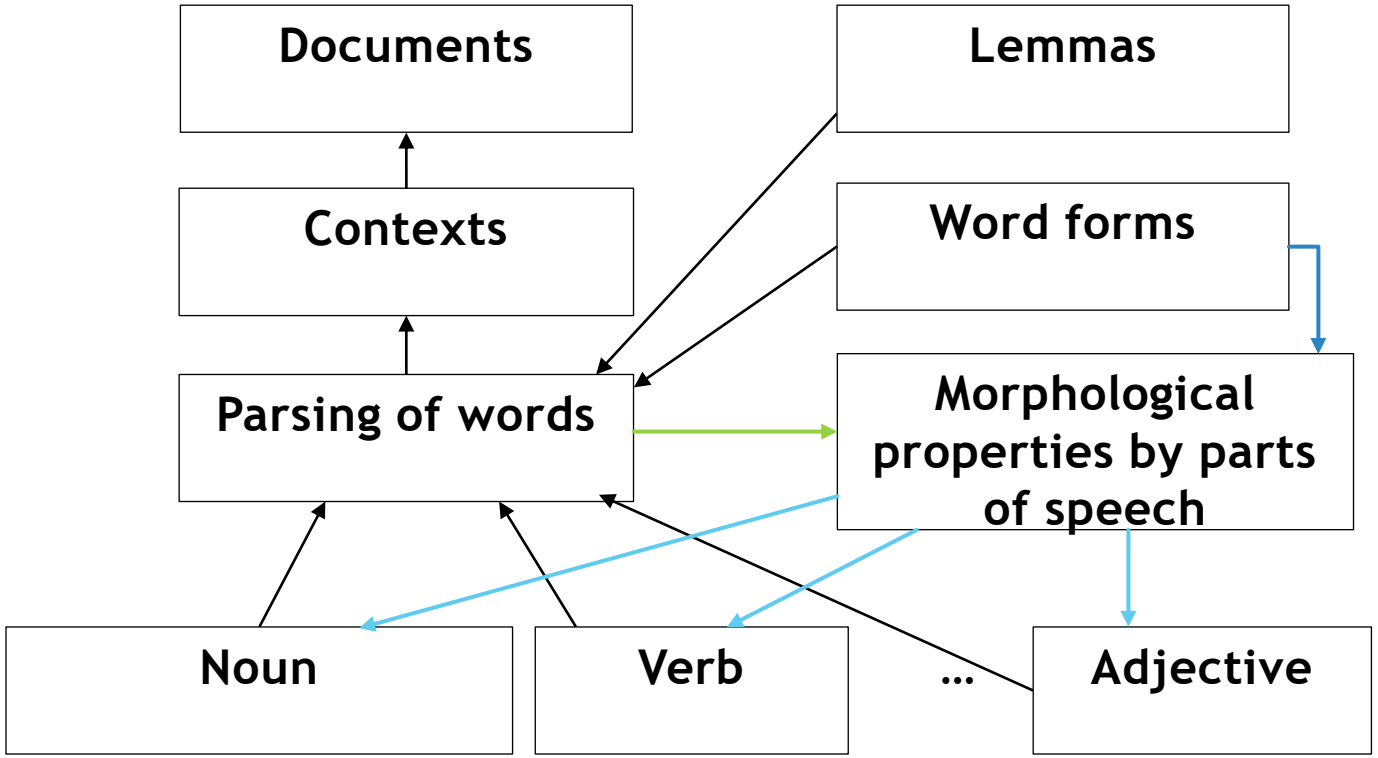
Damir Mukhamedshin

Tatarstan Republic Academy of Sciences
Institute of Applied Semiotics

Corpus Management System «Tugan Tel» (tugantel.tatar)

- Search of lexical and syntactic units
- Morphological, lexical and phrasal search
- Search the n-gram based on grammar
- Search using negative keywords and by parts of words
- Using arbitrary morphological formulas
- Execution of search queries in less than 1 second (MariaDB and Redis)
- Primarily aimed at supporting electronic corpora of Turkic languages

Relationships between Elements of the System



Building Search Queries

$$Q_1 = (\text{lemma}, \text{kitap}, \text{N|V}, \text{right}, 1, 10, \text{exact})$$
$$Q_2 = (\text{lemma}, \text{bar}, \text{V|ADV}, \text{right}, 1, 10, \text{exact})$$
$$Q = (Q_1, Q_2)$$

The Lexical Component of a Query

кита* (словоформа)= («китап», «китаплар», «китабы»…)

китап -“китаплар” (лемма) = («китап», «китапларны», «китабы»…, но не «китаплар»)

(словоформа, идентификатор) (2)

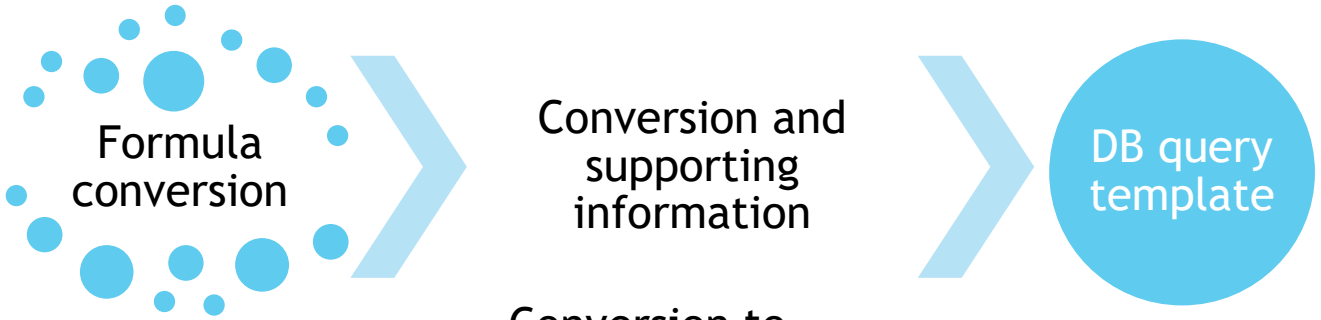
(лемма, идентификатор) (3)

$w \in W$ (4)

The Morphological Component of a Query

N, DIR, SG (5)
N|DIR|SG (6)
! (N|DIR|SG) (7)
! N|DIR, SG (8)

General Algorithm for Processing the Morphological Formulas



Arbitrary Formula Processing

(N|V,(!ADJ|POSS_2SG))

- (N|V,REPLACED_0)
- REPLACED_1

REPLACED_0

- REPLACED_S_0|POSS_2SG
- REPLACED_S_0|REPLACED_S_1
- **1..129, 140..149**

N|V,REPLACED_0

- RE…D_S_0|RE…D_S_1,RE…D_0
- REPLACED_S_2|REPLACED_S_3
- **120..129, 140..149**

An Example of a Query with the Logical Error

$$!(N|V), INF\_1$$

NOT (*noun* OR *verb*) AND *infinitive ending with -yrga*

Comparing Search Functionality in Different Corpus Management Systems

Corpus management system / Search query type	Sketch Engine (Czech corpus)	Yandex. Server (Russian corpus)	ExMA-RALDA	Tugan Tel Corpus Management System
Direct search	0.265 sec.	0.124 sec.	Local	0.117 sec.
Direct search by lemma	0.187 sec.	0.141 sec.	-	0.129 sec.
Direct search by wordform part	1.260 sec.	0.170 sec.	Local	0.591 sec.
Direct search by lemma part	-	0.188 sec.	-	0.517 sec.
Morphological search (AND)	0.272 sec.	1.170 sec.	-	0.073 sec.
Morphological search (OR)	22.570 sec.	13.320 sec.	-	0.117 sec.
Morphological search (NOT)	11.950 sec.	-	-	0.131 sec.
Morphological search (arbitrary formula with AND, OR, NOT)	14.120 sec.	-	-	0.063 sec.

Thank you for attention!

http://tugantel.tatar/

damirmuh@gmail.com

https://t.me/damirmuh