Collaboratively Developed Lexical Resources for Bulgarian with Application to Dictionaries and Reference Sources Compilation

Velislava Stoykova
Institute for Bulgarian Language – BAS
Bulgarian Academy of Sciences
vili1@bas.bg

Outline

- The survey of the experience at the Institute for Bulgarian Language - BAS in designing of available electronic text corpora and related language technology applications used for fast and up-to-date production of dictionaries and reference sources in Bulgarian language.

- The results of several past and on-going projects and their approaches to develop sustainable strategy to use the corpus-based approach both for academic research with theoretical value and for practical applications in compiling different types of dictionaries and reference sources are presented.

- The specific corpus-based applications for lexicography which use the balanced corpus content, the related specific software, and the representations of language-specific grammar features (like inflection, definite article, etc) are analysed.

Introduction

- The history and the use of Bulgarian National Corpus (BulNC) in compilation of recently published volumes of Dictionary of Bulgarian Language (DBL) (a multivolume academic explanatory dictionary).

- BulNC is a large-scale representative online available corpus of Bulgarian. It includes 320 000 000 words and incorporates modern original and translated Bulgarian texts from the middle of the XX century until the present.

- The BulNC incorporates four *general* sub-corpora: the Bulgarian Brown Corpus, the Structural Corpus of Bulgarian Electronic Documents 2001–2009, the Structural Corpus of Bulgarian Printed Editions 1945– 2009, and transcripts of spoken data.

BuINC

Structural Corpus of Bulgarian Electronic Documents

- Economy *0.8*
- Politics 2.2
- Law 2.6
- Medicine 0.7
- Sport 0.4
- Military 0.2
- Society 0.1
- Administration 4.6
- Journalism 7.3
- Fiction 9.1
- Other 1.2

BuINC

Structural Corpus of Bulgarian Printed Editions

- Prose 110,37
- Poetry 1,63
- Children's literature 3,82
- Plays 0,9
- Fables 0,06
- Folklore 3,28

- http://search.dcl.bas.bg/

BuINC

Resources for Specialized Lexica

- The corpus-based conceptual relation extraction has been used for the creation of highly structured and semantically-oriented up-to-date specialized lexical reference sources.

- The approach was developed within the framework of the innovative project “Conceptual Semantic Network Representation” aimed to use statistical approaches to extract the semantic conceptual relations from the specialized corpora in the domain of mathematics.

- The project uses web-based electronic text corpora in Bulgarian (MathWikiBul) and in English (MathWiki) consisting mostly of encyclopedic texts from Wikipedia (approx. 150 000 words each) in the domain of precalculus. It uses the comparative corpora analysis approach to relate the results for both languages.

Resources for Specialized Lexica

- The techniques for extraction of keywords, word concordances, collocations and co-occurrences. and the related statistical approaches incorporated in the Sketch Engine are used.

- Additionally, the comparison of the search results between the specialized and general corpora is used to extract the semantic conceptual relations.

  The constructed conceptual hierarchy of the basic domain concepts underlay the internal domain knowledge representation in ontology-like style. The resulting semantic interpretation is language-independent and present the domain knowledge.

Resources for Specialized Lexica

MathWiki: Extracted keywords

pic (1454) conic (88) rational (81) relation (54) trigonometric (141) functions (480) complex (289) constant (58) polynomial (210) finite (70) properties (75) expressed (52) theorem (89) function (635) plane (88) real (325) algebraic (62) converges (50) sequence (141) expression (51) multiplication (82) Main (80) corresponding (51) unit (72) cosine (80) mathematical (78) induction (60) length (93) matrices (162) linear (102) circle (80) elements (77) polynomials (114) formula (120) domain (79) argument (80) algebra (62) triangle (79) definition (81) operations (55) inverse (71) infinite (100) notion (70) terms (159) integers (61) coordinates (70) square (71) value (134) logarithm (118) graph (54) define (51) article (90) notation (81) convergence (62) theory (145) called (234) matrix (280) variables (75) values (103) series (273) sine (85) angles (83) equal (96) number (343) vector (226) angle (118) sin (65) form (180) equations (103) identities (52) element (77) natural (97) coefficients (71) variable (86) identity (66) term (102) tangent (53) numbers (401) negative (57) limit (68) calculus (51) defined (176) positive (91) sides (55) exponential (86) sum (91) product (135) set (341) vectors (104) polar (85) example (276) analysis (56) equation (154) zero (78) ordered (75) written (84) mathematics (106) coordinate (55) sets (65) general (99)

Resources for Specialized Lexica

- The basic conceptual relations definitions are generated by the use of collocations and co-occurrences of a related word.

- The search results from processing MathWikiBul and MathWiki for collocations of the keyword *function(s)* are similar for both Bulgarian and English. They present the most frequent content words which are most probably to be found with the keyword *function(s)* and define its semantically related concepts.

- The search results use *T−score* criterion for ranging the semantically related concepts but results according to *MI−score* and MI^3−score criteria are presented and used for the analysis.

Resources for Specialized Lexica

Corpus: MathWikiBul

Page 1 **of 13** Go

Next | Last

1.1 Изпъкнала	функция	1.2 Вдлъбната функция |
функция 1.2 Вдлъбната	функция	2 Свойства | |3 Вижте
дефинирана непрекъсната	функция	y(x), представена с крива
стойностите на тази	функция	съответно в точките x1 и x2.
за вдлъбната и изпъкнала	функция	се формира така: Изпъкнала
се формира така: Изпъкнала	функция	[редактиране] Функцията y(x
равенството. Вдлъбната	функция	[редактиране] Функцията y(x
: [pic]. Прието е линейната	функция	да бъде едновременно
част на графиката на	функция	, се наричат инфлексни точки

Resources for Specialized Lexica

Corpus: MathWiki

Page [1] of 33 [Go]

Next | Last

initiated by Descartes. A function , in mathematics, associates one
the argument of the function , also known as the input, with
quantity, the value of the function , also known as the output.
known as the output. A function assigns exactly one output to
given set. An example of a function is f(x) = 2x, a function which
a function is f(x) = 2x, a function which associates with every
f(5) = 10. The input to a function need not be a number, it can
object. For example, a function might associate the letter A with
describe or represent a function , such as a formula or algorithm

Resources for Specialized Lexica

Collocation candidates MathWikiBul

Page 1 Go

Next >

	Freq	T-score	MI	MI3
p/n аналитична	15	3.859	8.132	15.946
p/n комплексна	10	3.153	8.341	14.985
p/n реална	10	3.149	7.926	14.570
p/n непрекъсната	10	3.144	7.440	14.084
p/n тригонометрична	6	2.444	8.703	13.873
p/n примитивна	7	2.603	5.952	11.566
p/n холоморфна	4	1.989	7.466	11.466
p/n линейна	4	1.986	7.118	11.118
p/n обратна	4	1.984	6.926	10.926
p/n проста	3	1.724	7.703	10.873
p/n котангенс	3	1.724	7.703	10.873
p/n косинус	3	1.720	7.189	10.359

Resources for Specialized Lexica

Collocation candidates MathWiki

Page 1 Go

Next >		Freq	T-score	MI	MI3
p/n	exponential	65	8.001	7.040	19.085
p/n	inverse	29	5.309	6.152	15.868
p/n	rational	24	4.804	5.689	14.859
p/n	propositional	20	4.437	7.011	15.655
p/n	polynomial	23	4.544	4.253	13.301
p/n	complex	19	3.978	3.517	12.013
p/n	logarithm	12	3.268	4.146	11.316
p/n	trigonometric	11	3.072	3.764	10.683
p/n	increasing	8	2.794	6.357	12.357
p/n	tangent	8	2.721	4.716	10.716
p/n	relation	8	2.719	4.689	10.689
p/n	continuous	7	2.605	6.004	11.618

Specialized Lexica

The basic conceptual relations definitions are generated by the use of collocations and co-occurrences of a related word.

- The results present the concepts *exponential*, *rational*, *polynomial*, *complex*, *logarithmic*, *trigonometric*, etc as semantically related to the keyword *function(s)*.

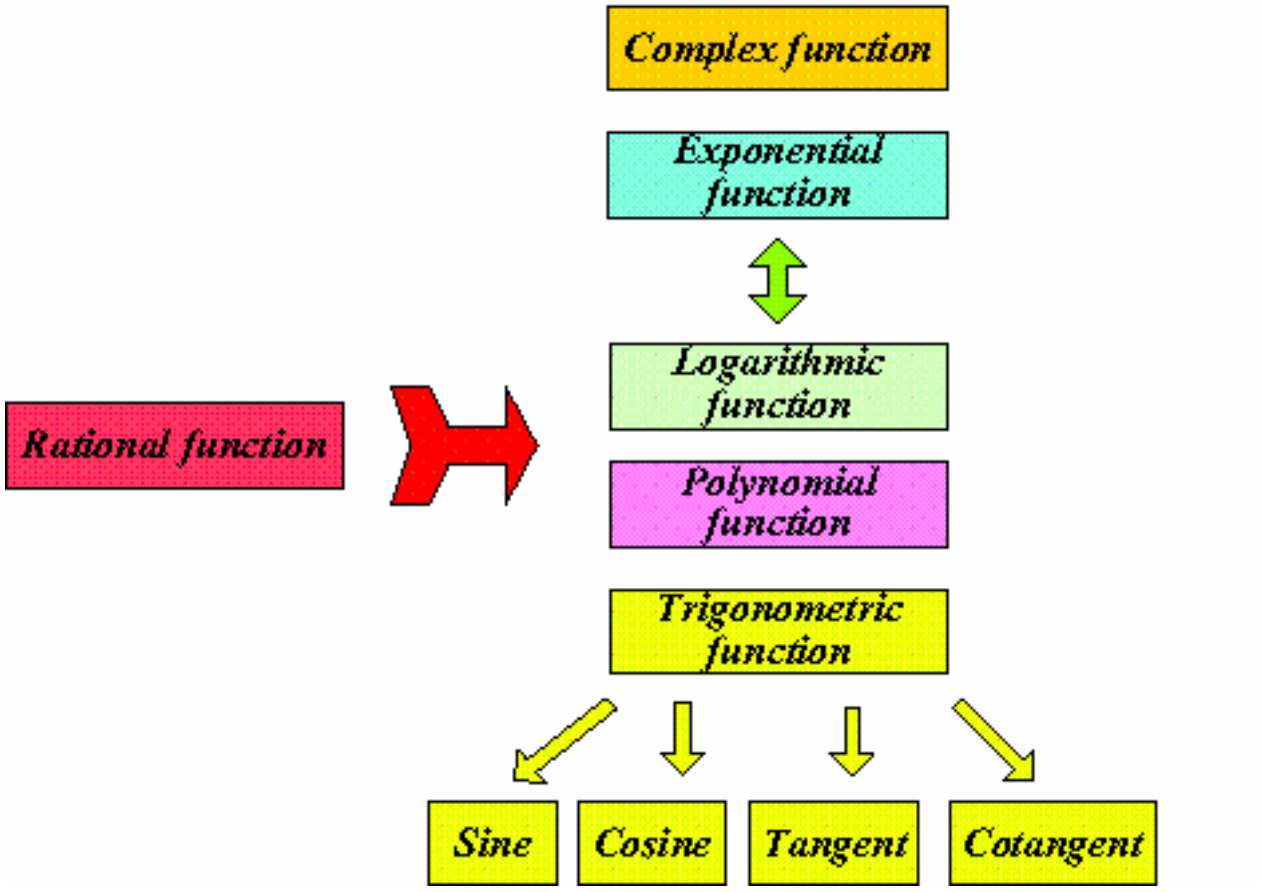
- However, the statistical similarity does not represent always synonymy and in this case, it expresses the hierarchical conceptual relations of the basic concept *function(s)* and the collocated concepts.

Semantic Conceptual Relations

- Generally, the conceptual semantic term relations extracted by collocations and co-occurrences mostly represent vertical semantic relations like hyponymy or hyperonymy.

- Thus, for our research results, we are using such interpretation and we define *polynomial function*, *exponential function*, and *rational function* as the most important hyponymic concepts of the very general hyperonym conceptual term *complex function*.

Semantic Conceptual Hierarchy



Conclusion

- The extracted terms and semantic relations show that statistically-based search technique for extracting keywords, collocations and co-occurrence words is effective approach for mathematical conceptual precalculus terms extraction from specialised web-corpus.

- The terms are evaluated on the base of their high frequency in the MathWikiBul and MathWiki corpora and their relatively low frequency in the BulNC and BNC. In fact, the term *precalculus* was not occurred in the BNC.

- The methodology is used for fast production of up-to-date terminological reference sources (like specialized dictionaries or thesauri) or building ontology (for defining the logical relations, conceptual relations or hierarchical semantic relations).

Thank you for your attention!