

Provides Fundamental Text Analysis for Asian, Middle Eastern, and European Languages

Full text search engines are ubiquitous. We access them daily on the Internet, in the office, and on our home computers. But, inside each search engine is sophisticated technology known as “computational linguistics” the automated analysis of digital text which enables it to be rapidly stored, searched, and retrieved.

Since 2000, the most widely used Internet and enterprise search engines have relied on Basis Technology's Rosette Base Linguistics (RBL) to provide essential linguistic services, including tokenization, lemmatization, decomposing, part-of-speech tagging, sentence boundary detection, and noun phrase extraction.

“Google selected Basis Technology to provide the Asian linguistic technology needed to create the ultimate Chinese, Japanese, and Korean search engine. This marks a key milestone in establishing Google as the preferred search engine for Internet users worldwide.”
 — Urs Hölzle, Fellow and Vice President, Google

The same linguistic technology which powers AOL, Ask, Google, Bing, and Yahoo is now available for search engines of all sizes and budgets. RBL is currently offered in 19 languages, with more under development:

Arabic	Chinese	Czech
Dutch	English	French
German	Greek	Hungarian
Italian	Japanese	Korean
Persian (Farsi/Dari)	Polish	Portuguese
Russian	Spanish	Urdu

RBL Approach

RBL relies on a morphological approach to analyzing text in different languages. This means that RBL works with the specific features of a given language: punctuation, actual words, word forms, and affixes. This yields far

more accurate results than statistical-based approaches, which can mistakenly produce non-words that result in false positive search results. In today's environment, where even tiny differences in accuracy can make or break the success of an application, such compromises are often unacceptable.

Key Features

RBL provides fundamental text analysis for Asian, Middle Eastern and European languages. Each language has its own linguistic complexity, and RBL handles them individually.

- **Segmentation/Tokenization** – determines the boundaries of the unique lexical tokens in input data, including locating punctuation and other special characters.
- **Lemmatization** – generates the dictionary base form for an inflected form of a verb or adjective.
- **Noun Decomposing** – breaks compound nouns into sub-compounds for accurate information retrieval.
- Generates the **linguistic stem** form of a word.
- Identifies a word's **part-of-speech** such as noun, verb or preposition.
- **Sentence Boundary Detection** – marks boundaries of individual sentences.
- **Context-Based Analysis** – analyzes forms not included in the main or custom dictionary based on context.
- **Base Noun Phrase Analysis** – identifies sets of words including a noun which describes a single nominal expression.
- Ignores user-identified **stop words**.

You may also be interested in...

- **Rosette Chinese Script Converter** – processes Chinese text and converts it to either Simplified or Traditional form, handling both the character variations and the word-level differences.
- **Rosette Japanese Orthographic Normalizer** – a dictionary-driven software that allows different orthographic forms of Japanese words to be converted to a standard canonical form.

Applications

RBL is a key component of the Rosette Linguistics Platform, and is ideal for any application which must process large volumes of multilingual text, including:

- Web Search
- Apache Lucene and Apache Solr
- E-Discovery and Digital Forensics
- Data Mining and Data Warehousing
- Enterprise Search
- Information Access Platforms
- Document and Media Exploitation
- Email and Instant Messaging

System Specifications

RBL is a component of the Rosette Linguistics Platform (RLP). It is portable and highly scalable, running on platforms ranging from laptop PCs to multi-CPU servers processing thousands of documents per second.

A fully-documented API is provided and may be accessed from applications written in C, C++, Java, and other languages. A command-line interface is also available for testing purposes.

Software Development Kits (SDK) are available for all major architectures, operating systems, and development environments.

For More Information

For more information or to request an evaluation copy, please write to info2009@basistech.com or call us at 617-386-2090.

The screenshot displays three paragraphs of Russian text with color-coded parts of speech. A legend on the right lists the parts of speech and their corresponding colors:

ADJ	adjective
ADV	adverb
CM	comma
CONJ	conjunction
DET	determiner
DIG	numerals (digits)
IREL	relative/interrogative pronoun
NOUN	common noun
PERS	personal pronoun
PREP	preposition
PRON	pronoun
PRONADV	pronominal adverb
PROP	proper noun
PTCL	particle
PUNCT	punctuation (other than CM or SENT)
SENT	sentence final punctuation
VAUX	auxiliary verb
VFIN	finite verb
VGER	verb gerund
VINF	verb infinitive
VPRT	verb participle

The text in the screenshot includes: "Украина и Евросоюз решили заключить соглашение об ассоциации, которое может быть подписано уже в 2009 году.", "Об этом во вторник в Париже заявил украинский президент Ющенко на совместной пресс-конференции с президентом Саркози, председателем Еврокомиссии Жозе Мануэлом Баррозу, комиссаром ЕС по вопросам внешней политики Хавьером Соланой.", "Совместная пресс-конференция европейских лидеров завершила саммит Евросоюз - Украина.", "Анализируя итоги встречи, Николя Саркози заявил украинской стороной условия безвизового режима.", "На пресс-конференции Саркози заявил, что в Европе насколько важен визовый вопрос для граждан Украины.", "Это амбициозная цель, которая займет время, но решили вместе работать", - сказал президент Франции Николя Саркози.

© 2009 Basis Technology Corporation. All rights reserved. "Basis Technology", "Rosette", and "We Put the World in the World Wide Web" are registered trademarks. All other company and product names mentioned are trademarks or registered trademarks of their respective owners. (2009-11-16)



Browse

www.basistech.com

Write

info2009@basistech.com

Call

617-386-2090 or 800-697-2062

Boston

One Alewife Center, Cambridge, MA 02140

San Francisco

282 Second Street, San Francisco, CA 94105

Washington, D.C.

13800 Coppermine Road, Herndon, VA 20171