

Rosette Base Linguistics

Search Multilingual Documents with High Accuracy

Rosette® Base Linguistics (RBL) enables information retrieval and text mining applications to process multilingual documents by providing essential linguistic services, including tokenization, lemmatization, and decomposing.

Each human language presents unique challenges, so RBL combines multiple technologies. For East Asian languages, proper tokenization is essential for accurate search results. RBL achieves this using morphological analysis, analyzing the specific features of a given language, such as punctuation, affixes, inflected words, and dictionary word forms. For European languages, RBL performs lemmatization which identifies the dictionary form of each word.

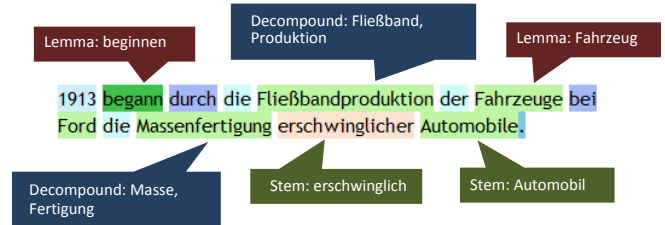
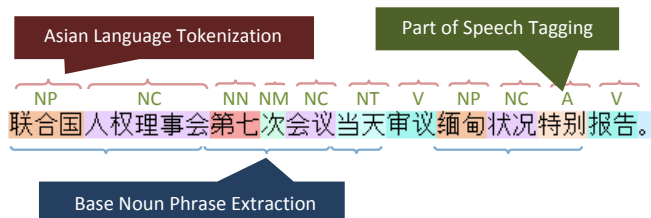
Simplistic methods, such as n-grams and stemming, deliver less accurate results. Through deep understanding of each language, we continually improve RBL with periodic dictionary updates and evaluate new approaches from the academic world.

“As the market expands internationally, one last significant market differentiator is the ability to mine text in foreign languages. Both Teragram and Inxight possessed this capability. But with those companies being acquired, Basis Technology is the main pure play in the market with multilingual text analysis capabilities.”

— Nick Patience, Research Director, Information Management, The 451 Group

BENEFITS

- **Broad language coverage** from a proven technology provider.
- **Features for search engines** include lemmatization, decomposing, and customizable dictionaries.
- **High-throughput and scalability** has made RBL the choice of major web and enterprise search engines.



KEY FEATURES

- **Tokenization** is a requirement for automated analysis of languages lacking spaces between words, such as Chinese, Japanese, and Korean.
- **Lemmatization** generates the dictionary form of each word, increasing search relevancy and slimming the search index—by indexing only lemmas (“cruise”) rather than all inflected forms (“cruising,” “cruised”).
- **Decomposing** breaks compound words into sub-components to increasing search relevancy for Germanic languages and Korean.
- **Part-of-Speech Tagging** is used during lemmatization to select the correct dictionary form of ambiguous words, such as the noun or verb “spoke.”
- **Sentence Boundary Detection** locates the start and end of sentences.
- **Noun Phrase Analysis** groups nouns and their modifiers, useful in document clustering and concept 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, Senior Vice President, Google

CUSTOMIZABLE FEATURES

Users can customize RBL for their data with these features:

- **User Dictionaries**, containing words or rules specific to your data, enable custom tokenization.
- The **Chinese Script Converter** enables Pan-Chinese search. Chinese speakers can search both simplified and traditional documents in one query, with results shown in the preferred script. This module achieves a high level of accuracy by converting word-by-word, rather than character-by-character.

- The **Japanese Orthographic Analyzer** is a dictionary-based normalizer that converts older, non-standard *kanji* to their modern forms, and normalizes spelling variants within *katakana*. Users can extend the normalization dictionary as needed.

SPOTLIGHT: ENGLISH

Linguistic analysis is not just for foreign languages; English also benefits. Lemmatization — finding the dictionary form of a word — broadens search to add *relevant* queries, in ways that traditional stemming cannot. This boosts recall without hurting precision.

Search Query	Traditional Stemming	Lemmatization using RBL	Comparison
animals	anim	animal	Two unrelated words may share a stem, in this case “anim”
animated	anim	animate	
several	sever	several	Stemming may have unintended consequences
children	children	child	Irregular verbs and nouns stump the stemmer
spoke	spoke	speak (when used as past tense verb) spoke (when used as noun)	

SPOTLIGHT: ARABIC

Arabic is a highly inflective language—affixes modify the start, end, and middle of words—such that exact string matching fails to locate many relevant hits. RBL increases search recall and precision via:

- **Normalization** standardizes spelling of words which vary due to a stylistic choice, careless spelling, or inconsistent use of diacritics.
- **Lemmatization** finds matches such as “two books” or “my books” when searching for “book”.



SPOTLIGHT: GERMANIC AND KOREAN

Danish, Dutch, German, Korean, Norwegian, and Swedish freely create compound words, which must be broken into subcomponents for indexing.

For example, in German, *Samstagmorgen* (“Saturday morning”) is a compound formed from *Samstag* (“Saturday”) and *Morgen* (“morning”). Decompounding *Samstagmorgen* enables matches to this word when searching for *Samstag* (“Saturday”).

SPOTLIGHT: CHINESE, JAPANESE, KOREAN, THAI

These widely-spoken languages are written without spaces between words. Morphological tokenization has many advantages over bigram or n-gram approaches, including reduction in index size and increase in search relevancy. Consider the problem of indexing [北京大学生物系 \(Beijing University Biology Department\)](#):

Traditional segmentation adds *six bigrams* to the index:

term position	1	2	3	4	5	6
bigrams	北京	京大	大学	学生	生物	物系

Bigram segmentation produces **non-words** and words which are **incorrect in this context**

Searching for 学生 (“student”) **incorrectly hits** “Beijing University Biology Department”.

Morphological tokenization adds *two properly segmented words* to the index:

term position	1	2
tokens	北京大学	生物系

Searching for 学生 (“student”) **correctly misses** “Beijing University Biology Department”.

AVAILABLE LANGUAGES

RBL is currently offered in these languages, with more under development:

Albanian	German	Polish
Arabic	Greek	Portuguese
Bulgarian	Hebrew	Romanian
Catalan	Hungarian	Russian
Chinese	Indonesian	Serbian
Croatian	Italian	Slovak
Czech	Japanese	Slovenian
Danish	Korean	Spanish
Dutch	Latvian	Swedish
English	Malay	Thai
Estonian	Norwegian	Turkish
Finnish	Pashto	Ukrainian
French	Persian	Urdu

SYSTEM PLATFORMS SUPPORTED

Software development kits (SDKs) and web services are available for the platforms listed below. Contact your sales representative for additional platform support.

AIX 6.1, PPC	Linux Ubuntu 10.x/11.x, IA32/AMD64
HP-UX 11i, IA64	MacOS
Linux CentOS 4.x/5.x, IA32/AMD64	Solaris 10, SPARC32/64, IA32/AMD64
Linux Debian 5.x, IA32/AMD64	Windows XP/Vista/7, IA32/AMD64
Linux Red Hat 4.x/5.x, IA32/AMD64	Windows Server 2003, 2008

VISIT www.basistech.com WRITE info2011@basistech.com CALL 617-386-2090

One Alewife Center
Cambridge, MA 02140

171 Second Street
San Francisco, CA 94105

2553 Dulles View Drive
Herndon, VA 20171

9-6 Nibancho, Chiyoda-ku
Tokyo 102-0084

