

# Rosette Language Identifier

## Automatically Classify Documents and Messages by Language and Encoding

Rosette® Language Identifier (RLI) scans text within documents to determine both the written language and character encoding scheme with very high accuracy. Identification of language and encoding is necessary for applications which categorize, search, process, and store text in any language.

RLI uses proprietary algorithms together with information-rich language profiles derived from hand-verified, training corpora for statistical analysis of all supported languages.

RLI can be used to determine:

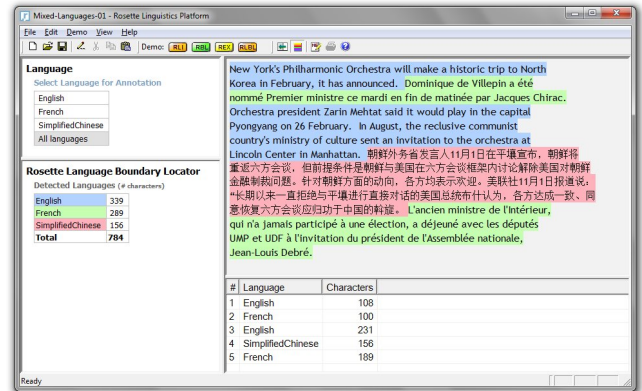
- The primary or dominant language of a document
- A complete list of all languages present in a multilingual document, and their percentage representation
- The start and end boundaries of each language region in a multilingual document—even if all the languages are written in the same script—such as English, French, German, or Italian
- The start and end boundaries of each writing system in a multiscript document, such as the Latin alphabet, Cyrillic alphabet, Japanese kana, or Chinese hanzi
- When languages have been transliterated or written with more than one alphabet, such as Arabic chat written in the Latin script

## APPLICATIONS

RLI is essential for any application which processes large volumes of multilingual text, including:

- Web and Enterprise Search Engines
- Information Access Platforms
- E-Discovery and Digital Forensics
- Document and Media Exploitation
- Data Mining and Data Warehousing
- E-mail and Instant Messaging

When used in conjunction with a transcoding engine (such as the Rosette Core Library for Unicode), RLI provides a powerful and immediate solution for converting a large, heterogeneous collection of text into a single, uniform representation based on the Unicode standard.



RLI's language boundary locator may be used to segment a multilingual document into monolingual regions.

## BENEFITS

Automatic language identification streamlines the processing of large quantities of text. Individual documents may be routed to language specialists, or automatically tagged for improved workflow. This process may also be combined with language-specific search engine plug-ins (such as Rosette Base Linguistics) to improve the quality of search results.

Although modern text encoding standards such as XML mandate the use of Unicode, many existing applications, documents, and data streams use "legacy encodings," such as ASCII, ISO 8859-1 (also known as "Latin 1"), Shift-JIS, and countless others.

RLI also provides a natural solution to the problem of migrating data repositories upward from older versions which support only legacy encodings to newer versions based on the Unicode standard.

## LANGUAGES AND ENCODINGS SUPPORTED

RLI supports 188 language/encoding pairs, covering 55 languages, 7 Latin script variants (transliterations), and 44 legacy encodings, plus the modern UTF-8 encoding for every language.

<b>Albanian</b> — ISO 8859-1, Windows-1252	<b>Lithuanian</b> — ISO 8859-13, Windows-1257
<b>Arabic</b> — ISO 8859-6, Windows-720, Windows-1256	<b>Macedonian</b> — ISO 8859-5, Windows-1251
<b>Arabic (transliterated)</b> — ISO 8859-1, Windows-1252, Windows-1256	<b>Malay</b> — ISO 8859-1, Windows-1252
<b>Bengali</b> — ISCII-Bengali	<b>Malayalam</b> — ISCII-Malayalam
<b>Bulgarian</b> — ISO 8859-5, Windows-1251, KOI8-R	<b>Norwegian</b> — ISO 8859-1, Windows-1252
<b>Catalan</b> — ISO 8859-1, Windows-1252	<b>Persian</b> — Windows-1256
<b>Chinese, Simplified</b> — GB-2312, HZ-GB-2312, GB-18030, ISO 2022-CN	<b>Persian (transliterated)</b> — ISO 8859-1, Windows-1252, Windows-1256
<b>Chinese, Traditional</b> — Big5, Big5-HKSCS	<b>Polish</b> — ISO 8859-2, Windows-1250
<b>Croatian</b> — Windows-1250	<b>Portuguese</b> — ISO 8859-1, Windows-1252
<b>Czech</b> — ISO 8859-2, Windows-1250	<b>Pushto</b> — ISO 8859-6, Windows-1256
<b>Danish</b> — ISO 8859-1, Windows-1252	<b>Pushto (transliterated)</b> — ISO 8859-1, Windows-1252
<b>Dutch</b> — ISO 8859-1, Windows-1252	<b>Romanian</b> — ISO 8859-2, Windows-1250
<b>English</b> — ISO 8859-1, Windows-1252	<b>Russian</b> — ISO 8859-5, Windows-1251, KOI8-R, IBM-866, Mac Cyrillic
<b>Estonian</b> — ISO 8859-13, Windows-1257	<b>Serbian</b> — ISO 8859-5, Windows-1251
<b>Finnish</b> — ISO 8859-1, Windows-1252	<b>Serbian (transliterated)</b> — ISO 8859-2, Windows-1250
<b>French</b> — ISO 8859-1	<b>Slovak</b> — Windows-1250
<b>German</b> — ISO 8859-1, Windows-1252	<b>Slovenian</b> — Windows-1250
<b>Greek</b> — ISO 8859-7, Windows-1253	<b>Somali</b> — ISO 8859-1, Windows-1252
<b>Gujarati</b> — ISCII-Gujarati	<b>Spanish</b> — ISO 8859-1, Windows-1252
<b>Hebrew</b> — ISO 8859-8, Windows-1255	<b>Swedish</b> — ISO 8859-1, Windows-1252
<b>Hindi</b> — ISCII-Hindi	<b>Tagalog</b> — ISO 8859-1, Windows-1252
<b>Hungarian</b> — ISO 8859-2, Windows-1250	<b>Tamil</b> — ISCII-Tamil
<b>Icelandic</b> — ISO 8859-1, Windows-1252	<b>Telugu</b> — ISCII-Telugu
<b>Indonesian</b> — ISO 8859-1, Windows-1252	<b>Thai</b> — Windows-874
<b>Italian</b> — ISO 8859-1, Windows-1252	<b>Turkish</b> — ISO 8859-9, Windows-1254
<b>Japanese</b> — EUC-JP, ISO-2022-JP, Shift-JIS, Shift-JIS-2004 (JIS X 0213)	<b>Ukrainian</b> — ISO 8859-5, Windows-1251, KOI8-R
<b>Kannada</b> — ISCII-Kannada	<b>Urdu</b> — ISO 8859-6, Windows-1256
<b>Korean</b> — EUC-KR, ISO-2022-KR	<b>Urdu (transliterated)</b> — ISO 8859-1, Windows-1252
<b>Kurdish</b> — Windows-1256	<b>Uzbek</b> — ISO 8859-5, Windows-1251, KOI8-R
<b>Kurdish (transliterated)</b> — ISO 8859-1, Windows-1252, Windows-1256	<b>Uzbek (transliterated)</b> — Windows-1251
<b>Latvian</b> — ISO 8859-13, Windows-1257	<b>Vietnamese</b> — TCVN, VIQR, VISCII, VNI, VPS

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 and scripting.

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

## EXPLORE FURTHER

For more information or to request an evaluation copy, please call us at 617-386-2090 or 800-697-2062, or write to [info2011@basistech.com](mailto:info2011@basistech.com). We will be happy to assist you in evaluating the performance of our product on your data.

VISIT [www.basistech.com](http://www.basistech.com) WRITE [info2011@basistech.com](mailto: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

