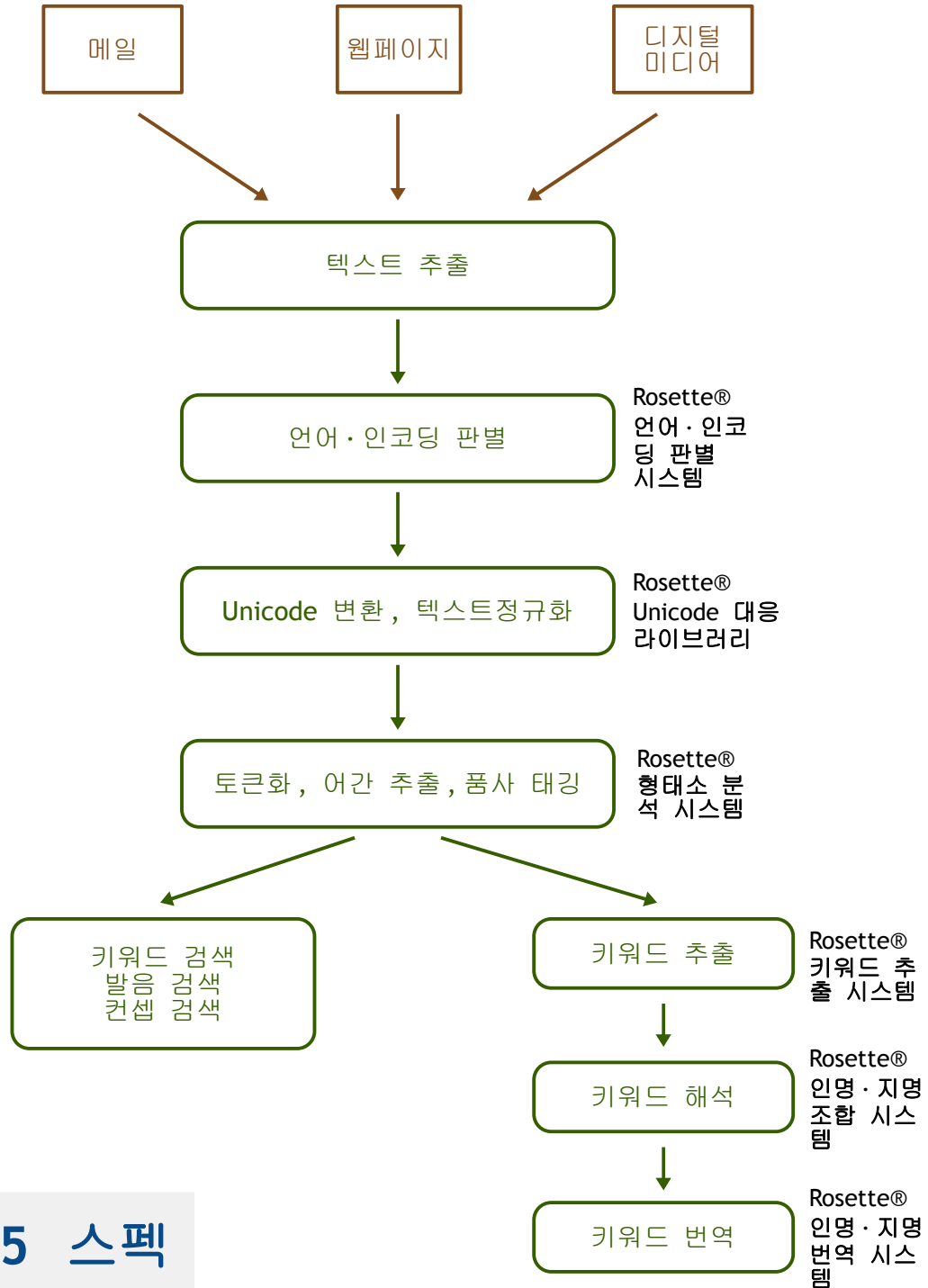


# Rosette 언어 처리 플랫폼



## 6.5 스펙

# Rosette 형태소 분석 시스템 (RBL)

	형태소 분석 시스템 (RBL)							REX	RNI	RNT
	토큰화	기본화	복합어분해	복사태깅	문장 경계 감지기	명사구 추출	로마자 라틴어/음성 읽기			
일본어	Y	Y	Y	Y	Y	Y	Y	N	N	
중국어 (간체자 + 번체자)	Y	n/a	n/a	Y	Y	Y	Y	Y	Y	
한국어	Y	Y	Y	Y	Y	Y	N	Y	Y	
아랍어	Y	Y	n/a	Y	Y	Y	Y	Y	Y	
파슈투어	N	N	N	N	N	N	N	P	Y	
페르시아어	Y	Y	n/a	N	Y	N	N	Y	Y	
우르두어	Y	Y	n/a	N	Y	N	N	Y	Y	
영어	Y	Y	n/a	Y	Y	Y	n/a	Y	Y	
프랑스어	Y	Y	n/a	Y	Y	Y	n/a	Y	N	
이탈리아어	Y	Y	n/a	Y	Y	Y	n/a	Y	N	
독일어	Y	Y	Y	Y	Y	Y	n/a	Y	N	
스페인어	Y	Y	n/a	Y	Y	Y	n/a	Y	N	
네델란드어	Y	Y	Y	Y	Y	Y	n/a	Y	N	
포르투갈어	Y	Y	n/a	Y	Y	Y	n/a	P	N	
러시아어	Y	Y	n/a	Y	Y	N	N	Y	N	
체코어	Y	Y	n/a	Y	Y	N	n/a	P	N	
히랍어	Y	Y	n/a	Y	Y	N	N	P	N	
폴란드어	Y	Y	n/a	Y	Y	N	n/a	P	N	
헝가리어	Y	Y	Y	Y	Y	N	n/a	P	N	

P — 부분 엔티티

n/a — 이 언어에는 사용 불가

# Rosette 키워드 추출 시스템

	사람	조직	장소	지정학 엔티티	시절	종교	국적	제목	신용카드 번호	거리	e-메일	위도/경도	행	숫자	ID번호	전화 번호	URL	UTM	날짜	시간	
	통계적인 / 가제티아 (각종 용어 사전)								정규 표현식												
일본어	S	S	S	S	S	S	S	S	R	R	R	R	R	R	R	R	R	R	R	R	R
중국어 (간체자 + 번체자)	S	S	S	S	S	S	S	S	R	R	R	R	R	R	R	R	R	R	R	R	R
한국어	S	S	S	S	S	--	--	S	R	R	R	R	R	--	R	R	R	R	R	R	R
아랍어	S	S	S	S	S	S	S	S	R	R	R	R	R	R	R	R	R	R	R	R	R
근대 페르시아어	S	S	S	S	S	S	S	G	R	R	R	R	R	R	R	R	R	R	R	R	R
우르두어	S	S	S	S	S	S	S	G	R	--	R	--	R	--	R	R	R	R	--	--	
영어	S	S	S	S	S	S	S	S	R	R	R	R	R	R	R	R	R	R	R	R	R
프랑스어	S	S	S	--	--	--	--	G	R	R	R	R	R	R	R	R	R	R	R	R	R
이탈리아어	S	S	S	--	--	--	--	G	R	R	R	R	R	R	R	R	R	R	R	R	R
독일어	S	S	S	--	--	--	--	G	R	R	R	R	R	R	R	R	R	R	R	R	R
스페인어	S	S	S	--	--	--	--	G	R	R	R	R	R	R	R	R	R	R	R	R	R
네델란드어	S	S	S	--	--	--	--	G	R	R	R	R	R	R	R	R	R	R	R	R	R
포르투갈어	--	--	--	--	--	--	--	G	R	R	R	R	R	R	R	R	R	R	R	R	R
헝가리어	--	--	--	--	--	--	--	--	R	--	R	R	R	--	R	R	R	R	R	R	--
체코어	--	--	--	--	--	--	--	--	R	--	R	R	R	--	R	R	R	R	R	R	--
희랍어	--	--	--	--	--	--	--	--	R	--	R	R	R	--	R	R	R	R	R	R	--
러시아어	S	S	S	S	S	G	G	S	R	R	R	R	R	R	R	R	R	R	R	R	R
폴란드어	--	--	--	--	--	--	--	--	R	--	R	R	R	--	R	R	R	R	R	R	--

- S 통계적인
- G 가제티아 (각종 용어 사전)
- R 정규 표현식 법칙

# Rosette 언어·인코딩 판별 시스템 (RLI)

합계 언어: **55, plus 7 Latin-script variants**  
 전체 프로파일: **191**  
 합계 인코딩: **39**

언어	지원 인코딩	인코딩 수
알바니아어	Windows-1252, ISO-8859-1	3
아랍어 (Arabic script)	Windows-1256, Windows-720, ISO-8859-6	4
아랍어 (Latin script)	Windows-1252, ISO-8859-1, Windows-1256	4
벵골어	ISCII-Bengali	2
불가리아어	Windows-1251, KOI8-R, ISO-8859-5	4
카탈로니아어	Windows-1252, ISO-8859-1	3
중국어간체자	HZ-GB-2312, GB18030, GB2312, ISO-2022-CN	5
중국어번체자	Big5	2
크로아티아어	Windows-1250, ISO-8859-2	3
체코어	Windows-1250, ISO-8859-2	3
덴마크어	Windows-1252, ISO-8859-1	3
네델란드어	Windows-1252, ISO-8859-1	3
영어	Windows-1252, ISO-8859-1	3
에스토니아어	Windows-1257, ISO-8859-13	3
핀란드어	Windows-1252, ISO-8859-1	3
프랑스어	Windows-1252, ISO-8859-1	3
독일어	Windows-1252, ISO-8859-1	3
희랍어	Windows-1253, ISO-8859-7	3
그자라티어	ISCII-Gujarati	2
헤브라이어	Windows-1255, ISO-8859-8	3
힌디어	ISCII-Devanagari	2
헝가리어	Windows-1250, ISO-8859-2	3
아이슬란드어	Windows-1252, ISO-8859-1	3
인도네시아어	Windows-1252, ISO-8859-1	3
이탈리아어	Windows-1252, ISO-8859-1	3
일본어	EUC-JP, Shift_JIS-2004, Shift_JIS, ISO-2022-JP	5
칸나다어	ISCII-Kannada	2
한국어	EUC-KR, ISO-2022-KR	3
쿠르드어 (Arabic script)	Windows-1256	2
쿠르드어 (Latin script)	Windows-1252, ISO-8859-1, Windows-1256	4
라트비아어	Windows-1257, ISO-8859-13	3
리투아니아어	Windows-1257, ISO-8859-13	3
마케도니아어	Windows-1251, ISO8859-5	3
말레이어	Windows-1252, ISO-8859-1	3
말라얄람 어	ISCII-Malayalam	2
노르웨이어	Windows-1252, ISO-8859-1	3
파슈투어 (Arabic script)	Windows-1256	2
파슈투어 (Latin script)	Windows-1252, ISO-8859-1, Windows-1256	4
페르시아어 (Arabic script)	Windows-1256	2
페르시아어 (Latin script)	Windows-1252, ISO-8859-1, Windows-1256	4
폴란드어	Windows-1250, ISO-8859-2	3
포르투갈어	Windows-1252, ISO-8859-1	3
루마니아어	Windows-1250, ISO-8859-2	3
러시아어	Windows-1251, KOI8-R, ISO-8859-5, IBM866, x-mac-cyrillic	6
세르비아어 (Cyrillic script)	Windows-1251, ISO-8859-5	3
세르비아어 (Latin script)	Windows-1250, ISO-8859-2	3
슬로바키아어	Windows-1250, ISO-8859-2	3
슬로베니아어	Windows-1250, ISO-8859-2	3
소말리어	Windows-1252, ISO-8859-1	3
스페인어	Windows-1252, ISO-8859-1	3
스웨덴어	Windows-1252, ISO-8859-1	3
타갈로그어	Windows-1252, ISO-8859-1	3
타밀어	ISCII-Tamil	2
테로그어	ISCII-Telugu	2
타이어	Windows-874	2
타키어	Windows-1254, ISO-8859-9	3
우크라이나어	Windows-1251, KOI8-R, ISO-8859-5	4
우르두어 (Arabic script)	Windows-1256	2
우르두어 (Latin script)	Windows-1252, ISO-8859-1, Windows-1256	4
우즈베키스탄어 (Cyrillic script)	Windows-1251, KOI8-R, ISO-8859-5	4
음역 우즈베키스탄어	Windows-1251	2
베트남어	TCVN, VIQR, VISCII, VPS, VNI	6

인코딩 리스트
Big5
EUC-JP
EUC-KR
GB18030
GB2312
HZ-GB-2312
IBM866
ISCII-Bengali
ISCII-Devanagari
ISCII-Gujarati
ISCII-Kannada
ISCII-Malayalam
ISCII-Tamil
ISCII-Telugu
ISO-2022-CN
ISO-2022-JP
ISO-2022-KR
ISO-8859-5
ISO-8859-6
ISO-8859-7
ISO-8859-8
ISO-8859-9
ISO-8859-13
Shift_JIS
Shift_JIS-2004
TCVN
VIQR
VISCII
VNI
VPS
Windows-1251
Windows-1252
Windows-1253
Windows-1254
Windows-1255
Windows-1256
Windows-720
Windows-874
x-mac-cyrillic

## 하드웨어 / OS

OS	버전	건축물	컴파일러	버전
AIX	5.2	PowerPC	Visual Age	6.x
FreeBSD	4.8	IA32	gcc	3.4.4
FreeBSD	6.0	IA32	gcc	3.4.4
FreeBSD	6.0	AMD64	gcc	3.4.4, 3.4.5
HP-UX	11.00	PA-RISC32	aCC	3.33
HP-UX	11.22	IA64	aCC	5.41
Linux Red Hat	ES 2.1	IA32	gcc	3.2
Linux Red Hat	ES 3.0	IA32	gcc	3.2; 3.4; 4.0
Linux Red Hat	ES 3.0	AMD64	gcc	3.4; 4.0
Linux Red Hat	ES 4.0	IA32	gcc	3.2; 3.4; 4.0
Linux Red Hat	ES 4.0	AMD64	gcc	3.4; 4.0
Linux	Fedora Core 4	IA32	gcc	3.2; 3.4; 4.0
Linux	Fedora Core 4	AMD64	gcc	3.4; 4.0
Linux	Fedora Core 5	IA32	gcc	3.2; 3.4; 4.0
Linux	Fedora Core 5	AMD64	gcc	4.1
Linux	Debian 3.1	IA32	gcc	3.2; 3.4; 4.0
Linux	Debian 3.1	AMD64	gcc	3.4; 4.0
Linux	Suse EL 10	IA32	gcc	3.2; 3.4; 4.0
Linux	Suse EL 10	AMD64	gcc	4.1
Mac OS	10.5	IA32/IA64	gcc	4.0
Solaris	8	SPARC32	Forte Developer CC	5.2
Solaris	8	SPARC64	Forte Developer CC	5.2
Solaris	9	SPARC32	Forte Developer CC	5.8
Solaris	9	SPARC32	gcc	3.4.5
Solaris	9	SPARC64	Forte Developer CC	5.8
Solaris	9	IA32	gcc	4.1
Solaris	10	SPARC32	Forte Developer CC	5.8
Solaris	10	SPARC64	gcc	4.1.2
Solaris	10	SPARC64	Forte Developer CC	5.8
Solaris	10	IA32	Forte Developer CC	5.8
Solaris	10	IA32	gcc	3.4
Solaris	10	AMD64	Forte Developer CC	5.8
Solaris	10	AMD64	gcc	4.1.2
Windows	NT, XP, 2003	IA32	Visual C++	7.1; 8.0
Windows	2003	AMD64	Visual C++	8.0
Windows	Vista	IA32	Visual C++	7.1; 8.0
Windows	Vista	AMD64	Visual C++	8.0

### SDK Distribution

C++ platform  
 DLL implementation  
 Java JNI and native Java APIs  
 .NET (Windows)

## About Basis Technology

Basis Technology is the leading provider of software solutions for extracting meaningful intelligence from unstructured multilingual text.

Our products and services are used by over 250 major firms, including Cisco, EMC, Endeca, HP, Microsoft, Oracle, and Symantec.

Our text analysis products are widely used in the U.S. defense and intelligence industry by such firms as CACI, Lockheed Martin, MITRE, Northrop Grumman, SAIC, and SRI. We are also the top provider of multilingual search technology to web search engines, such as AOL, Ask.com, Google, Windows Live, and Yahoo.

Our Rosette Linguistics Platform is the world's most widely-used family of commercial software products for multilingual text retrieval and analysis. Rosette provides such services as automatic language identification, Unicode text normalization, and entity extraction from unstructured text.

"Basis Technology" and "Rosette" are registered trademarks of Basis Technology Corporation. All other company and product names mentioned herein are trademarks or registered trademarks of their respective owners.

© 2009 Basis Technology Corporation. All rights reserved.  
(AD09.10.20)

## Selected U.S. Government Customers

CACI International  
Hitachi  
In-Q-Tel  
Lockheed Martin  
MITRE  
NEC  
Northrop Grumman  
SAIC  
SRI  
U.S. Department of Defense  
U.S. Department of Justice  
U.S. Intelligence Community

## Selected Commercial Customers

Attivio  
Autodesk  
Cisco  
EMC  
Endeca  
Google  
HP  
Kroll  
Mark Logic  
Microsoft  
Oracle  
Symantec  
Yahoo!  
WiseNut



### URL

[www.basistech.co.kr](http://www.basistech.co.kr)

### E-mail

[info@basistech.jp](mailto:info@basistech.jp)

### Phone

+81-(0)3-3511-2947

### 일본 도쿄

9-6 Niban-cho, Chiyoda-ku, Tokyo 102-0084 Japan

### 미국 본사 (보스턴)

One Alewife Center, Cambridge, MA 02140 USA

### 워싱턴 D.C. 지사

13800 Coppermine Road, Herndon, VA 20171 USA