SUPPORTED LANGUAGES – CV23

Table of Contents

Detected Languages

The following table lists exhaustively all the languages detected by CloudView. The Tokenized Languages section lists the languages which support semantic treatment.
Note: The Language Detector document processor used by CloudView for language detection is more accurate for some languages than others.
For example, it detects correctly English, Russian, Greek, Hebrew, and Japanese most of the time.
Note that the more consecutive words in a specific language the document has, the more precision you get.

A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**	**L**
Abkhazian	Bashkir	Catalan	Danish	English	Faroese	Galician	Hausa	Icelandic	Japanese	Kalaallisut	Lao
Afar	Basque	Central Khmer	Dutch	Esperanto	Fijian	Georgian	Hebrew	Indonesian	Java	Kannada	Latin
Afrikaans	Belarusian	Chinese	Dzongkha	Estonian	Finnish	German	Hindi	Interlingua		Kashmiri	Latvian
Albanian	Bengali	Corsican			French	Greek	Hungarian	Interlingue		Kazakh	Lingala
Amharic	Bihari languages	Croatian				Guarani		Inuktitut		Kinyarwanda	Lithuanian
Arabic	Bislama	Czech				Gujarati		Inupiaq		Kirghiz	
Armenian	Breton							Irish		Korean	
Assamese	Bulgarian							Italian		Kurdish	
Aymara	Burmese										
Azerbaijani											

M	**N**	**O**	**P**	**Q**	**R**	**S**	**T**	**U**	**V**	**W**	**X**	**Y**	**Z**
Macedonian	Nauru	Occitan	Persian	Quechua	Romanian	Samoan	Tagalog	Ukrainian	Vietnamese	Welsh	Xhosa	Yiddish	Zhuang
Malagasy	Nepali	Oriya	Polish		Romansh	Sango	Tajik	Urdu	Volapük	Western Frisian		Yoruba	Zulu
Malay	Norwegian	Oromo	Portuguese		Rundi	Sanskrit	Tamil	Uyghur		Wolof			
Malayalam			Punjabi		Russian	Scottish Gaelic	Tatar	Uzbek					
Maltese			Pushto			Serbian	Telugu						
Maori						Serbo-Croatian	Thai						
Marathi						Shona	Tibetan						

Moldavian
Mongolian

Sindhi
Sinhala
Slovak
Slovenian
Somali
Sotho, Southern
Spanish
Sundanese
Swahili
Swati
Swedish

Tigrinya
Tonga
Tsonga
Tswana
Turkish
Turkmen

Twi

Tokenized Languages

Cloudview tokenizes many languages by default (known as Standard support), and provides additional semantic features.
To get additional semantic features, you can purchase the Extended Languages add-on.
Use the table below to decide whether to analyze the languages in your corpus with standard or Extended Languages (Basis Tech) tokenizers.
For details on setting up tokenizers, see the CloudView Configuration guide > Configuring Data Processing > Tokenizing Text section.
The CloudView Configuration guide also describes the various semantic features available in CloudView.

Flag	Supported by
Std	Standard tokenizer
Jap	Specific CloudView tokenizer. For more advanced semantic treatment, move to *Extended Languages*.
Zhi	Specific CloudView tokenizer. For more advanced semantic treatment, move to *Extended Languages*.
Ext	*Extended Languages* (Basis Tech) tokenizers. These tokenizers are embedded in CloudView and require the *Extended Languages* license.
Ext+	*Extended Languages* (Basis Tech) additional tokenizers (open source add-on). These tokenizers are not embedded in CloudView and require the *Extended Languages* license. This add-on is available from V6R2016x.R4. Download the add-on at http://www.basistech.com/language-download/ and unzip its content to <INSTALLDIR>/resource/all-arch.
-	not supported

	Tokenization & Normalization	Lemmatization	Stemming	De - agglutination	Sentence Boundary Detection	Part of speech	Noun Phrase Detection	Named Entity Extraction
Afrikaans	**Std**	-	-	-	-	-	-	-
Albanian	**Std**/ **Ext+**	**Ext+**	-	-	**Ext+**	**Std**	-	-
Arabic	**Std**/ **Ext**	**Ext**	-	-	**Ext**	**Ext**	**Ext**	**Ext**
Basque	**Std**	-	-	-	-	-	-	-
Bengali	**Std**	-	-	-	-	-	-	-
Breton	**Std**	-	-	-	-	-	-	-
Bulgarian	**Std**/ **Ext+**	**Ext+**	-	-	**Ext+**	**Std**	-	-
Catalan	**Std**/ **Ext+**	**Ext+**	-	-	**Ext+**	**Std**	-	-
Chinese	**Zhi**/ **Ext**	-	-	**Ext**	**Ext**	**Ext**	**Zhi**/ **Ext**	**Ext**
Croatian	**Std/ Ext+**	**Ext+**	-	-	-	**Std**	-	-
Czech	**Std**/ **Ext**	**Ext**	-	-	**Std**/ **Ext**	**Ext**	-	**Ext** (only for dates)
Danish	**Std**/ **Ext**	**Ext**	**Std**	**Ext**	**Std**/ **Ext**	**Std**	**Std**	-
Dutch	**Std**/ **Ext**	**Ext**	**Std**	**Std**/ **Ext**	**Std**/ **Ext**	**Ext**	**Std**/ **Ext**	**Std**/ **Ext**
English	**Std**/ **Ext**	**Std**/ **Ext**	**Std**	-	**Std**/ **Ext**	**Std**/ **Ext**	**Std**/ **Ext**	**Std**/ **Ext**
Estonian	**Std/ Ext+**	**Ext+**	-	-	**Std**/ **Ext+**	**Std**	-	-
Farsi (Persian)	**Ext**	-	-	-	**Ext**	-	-	**Ext**

Finnish	**Std/ Ext**	-	-	-	**Std/ Ext**	**Std**	-	-
French	**Std/ Ext**	**Std/ Ext**	**Std**	-	**Std/ Ext**	**Std/ Ext**	**Std/ Ext**	**Std/ Ext**
German	**Std/ Ext**	**Std/ Ext**	**Std**	**Std/ Ext**	**Std/ Ext**	**Ext**	**Std/ Ext**	**Std/ Ext**
Greek	**Std/ Ext**	**Ext**	-	-	**Ext**	**Ext**	-	**Ext** (except places)
Hebrew	**Std/ Ext**	-	-	-	**Ext**	**Std**	-	**Ext**
Hindi	**Std**	-	-	-	-	-	-	-
Hungarian	**Std/ Ext**	**Ext**	-	**Ext**	**Std/ Ext**	**Ext**	-	**Ext** (except places)
Indonesian	**Std/ Ext+**	**Ext+**	-	-	**Ext+**	**Std**	-	-
Italian	**Std/ Ext**	**Std**	**Std**	-	**Std/ Ext**	**Std/ Ext**	**Std/ Ext**	**Std/ Ext**
Japanese	**Jap/ Ext**	**Ext**	-	**Ext**	**Ext**	**Ext**	**Jap/ Ext**	**Ext**
Korean	**Ext**	**Ext**	-	**Ext**	**Ext**	**Ext**	-	**Ext**
Latin	**Std**	-	-	-	-	-	-	-
Latvian	**Std/ Ext+**	**Ext+**	-	-	**Ext+**	**Std**	-	-
Malay	**Std/ Ext+**	**Ext+**	-	-	**Ext+**	**Std**	-	-
Nepali	**Std**	-	-	-	-	-	-	-
Norwegian	**Std/ Ext**	**Ext**	-	**Std/ Ext**	**Std/ Ext**	**Std**	**Std**	-
Pashto	**Ext**	-	-	-	**Ext**	-	-	**Ext**
Polish	**Std/ Ext**	**Ext**	-	-	**Std/ Ext**	**Ext**	**Std**	**Ext** (except places)

Portuguese	**Std/ Ext**	**Std/ Ext**	**Std**	-	**Std/ Ext**	**Ext**	**Std/ Ext**	**Std/ Ext**
Romanian	**Std/ Ext**	**Ext**	**Std**	-	**Std/ Ext**	**Std**	**Std**	-
Russian	**Std/ Ext**	**Std/ Ext**	**Std**	-	**Std/ Ext**	**Ext**	**Std**	**Ext** (except places)
Serbian	**Std/ Ext+**	**Ext+**	-	-	-	**Std**	-	-
Slovak	**Std/ Ext+**	**Ext+**	-	-	**Std/ Ext+**	**Std**	**Std**	-
Slovenian	**Std/ Ext**	**Ext+**	-	-	**Std/ Ext+**	**Std**	**Std**	-
Spanish	**Std/ Ext**	**Std/ Ext**	**Std**	-	**Std/ Ext**	**Ext**	**Std/ Ext**	**Std/ Ext**
Swedish	**Std/ Ext**	**Ext**	**Std**	**Ext**	**Std/ Ext**	**Std**	**Std**	-
Thai	**Ext**	-	-	-	-	-	-	-
Turkish	**Std/ Ext**	-	**Std**	-	**Std/ Ext**	**Std**	-	-
Ukrainian	**Std/ Ext+**	**Ext+**	-	-	**Ext+**	**Std**	-	-
Urdu	**Ext**	-	-	-	**Ext**	-	-	**Ext**
Welsh	**Std**	-	-	-	-	-	-	-