Bitext Linguistic Services

Bitext offers a multilingual platform to analyze & tag text at two levels:

- lexical – word-level
- syntactic and semantic – sentence-level

The platform has been designed for enterprise use: one single API for any language, platform independent (runs on Linux, Windows, Android, iOS), maximum performance/throughput, minimum footprint and easy scalability. As for deployment, customized solutions are available both for cloud or on-premise.

Lexical Level and Lemmatization

At the lexical level, the main component is the lemmatizer, integrated with tools for decompounding or word segmentation (required by some languages to perform proper lemmatization).

The lemmatizer can be additionally packaged to cover a language analysis full pipeline, from sentence segmentation to full parsing, including tools like spell checking.

Both components of the lemmatizer, data and software, can be distributed integrated or separately. All these tools are available in 78 languages and 27 language variants.

Syntactic Level and Parsing

At the syntactic level, the parser is the main component. The parser analyzes the structure of the sentences in the text and is used for tasks like POS Tagging and Phrase Extraction. Additionally, it is used as the base component for various semantic-level tasks like Entity Extraction, Topic-Level Sentiment Analysis or Categorization. We have developed parsers for 21 languages and are always adding new languages.

Linguistic Services Overview

Lexical Services (No Grammar)

Language identification	Detect the language(s) used in each sentence of a longer input text Applicable to all languages Example: Oui! I love Paris → "Oui!" – French, "I love Paris" – English
Sentence segmentation	Splits text into sentences, according to language-specific punctuation rules. Applicable to all languages. Example: Hello! How are you doing? → Hello! | How are you doing?
Tokenization	Splits a sentence into words, according to language-specific space and punctuation rules. Applicable to most languages (except Chinese, Japanese, Vietnamese, Thai…) Example: How are you doing? → How | are | you | doing | ?
Word segmentation (no-space tokenization)	Split text into words for languages that do not use spaces to separate them. Applicable to Chinese, Japanese, Vietnamese, Thai… Example: 把音量调低一点→ 把 | 音量 | 调低 | 一点
Lemmatization	Return the possible roots for a word form Applicable to most languages (except Chinese, Vietnamese, Thai…) Example: running → run
Decompounding	Split compound words/tokens into its individual component words. Applicable to German, Dutch, Norwegian, Swedish, Korean… Example: Rindfleischetikettierung → Rind | Fleisch | Etikettierung
Spelling	Check if a word is spelled correctly Applicable to all languages Example: excelent → incorrect

Syntactic and Semantic Services (Grammar and Meaning)

POS Tagging	Return the parts of speech for each word in a sentence Applicable to all languages Example: He runs back home → “He” – pronoun, “runs” – verb, “back” – preposition, “home” - noun
Phrase Extraction	Returns the constituents (noun phrases, verb phrases…) of a sentence Applicable to all languages Example: John’s sister was performing in the theatre → “John’s sister” – NP, “was performing” – VP, “in the theatre” – PP
Parsing	Produce a tree with the hierarchical constituent parts of a sentence (words, phrases, clauses…) Applicable to all languages
Entity Extraction	Detect proper names (people, places…) and other special text (phones, URLs…) Applicable to all languages Example: John lives in New York → “John” – person name, “New York” – place
Topic-Based Sentiment Analysis	Returns the sentiment and corresponding topic of opinions in text Applicable to all languages Example: I hate my old phone → opinion: “hate” (negative), topic: “my old phone”
Topic Detection	Returns topics of opinions in text Applicable to all languages Example: I hate my old phone → topic: “my old phone”
Categorization	Returns the categories applicable to a text, based on pre-defined rules Applicable to all languages Example: John is feeling great. → HAPPINESS [RULE: feel + great → HAPPINESS] Example: John was weeping like a willow. → SADNESS [RULE: weep + like + willow → SADNESS]

Other

Bots & Assistants	Evaluation and training data generation for chatbots/assistants

Linguistic Services Matrix - Languages

LANGUAGE	ISO	TIER	LANGID	SENTENCE SEG	TOKENIZATION	WORD SEG	LEMMATIZATION	DECOMPOUNDING	SPELLING	POS TAGGING	PHRASE	PARSE	ENTITIES	SENTIMENT	CATEGORIZATION	TOPIC	BOTS & ASSISTANTS
Afrikaans	AF	2	x	x	x		x		x								
Albanian	SQ	2	x	x	x		x		x								
Amharic	AM	3	x	x	x		x		x								
Arabic	AR	3	x	x	x		x		x	x			x				
Armenian	HY	3	x	x	x		x		x								
Assamese	AS	3	x	x	x		x		x								
Azeri	AZ	3	x	x	x		x		x								
Basque	EU	3	x	x	x		x		x								
Belarusian	BE	2	x	x	x		x		x								
Bengali	BN	3	x	x	x		x		x								
Bulgarian	BG	2	x	x	x		x		x								
Burmese	MY	3	x	x	x		x		x								
Catalan	CA	1	x	x	x		x		x	x	x	x		x		x	
Chinese	ZH	3	x	x		x	x		x	x		x	x	x		x	
Croatian	HR	2	x	x	x		x		x			x					
Czech	CS	2	x	x	x		x		x	x		x					
Danish	DA	1	x	x	x		x		x	x		x					x
Dutch	NL	1	x	x	x		x	x	x	x	x	x	x	x	x	x	x
English	EN	1	x	x	x		x		x	x	x	x	x	x	x	x	x
Esperanto	EO	2	x	x	x		x		x								
Estonian	ET	3	x	x	x		x		x								

Finnish	FI	3	x	x	x		x		x								
French	FR	1	x	x	x		x		x	x	x	x	x	x	x	x	x
Galician	GL	1	x	x	x		x		x								
Georgian	KA	3	x	x	x		x		x								
German	DE	1	x	x	x		x	x	x	x	x	x	x	x	x	x	x
Greek	EL	2	x	x	x		x		x								
Gujarati	GU	3	x	x	x		x		x								
Hebrew	HE	3	x	x	x		x		x								
Hindi	HI	3	x	x	x		x		x								
Hungarian	HU	3	x	x	x		x		x	x		x					
Icelandic	IS	2	x	x	x		x		x								
Indonesian	ID	2	x	x	x		x		x								
Irish Gaelic	GA	2	x	x	x		x		x								
Italian	IT	1	x	x	x		x		x	x	x	x	x	x	x	x	x
Japanese	JP	3	x	x		x	x		x	x			x				
Kannada	KN	3	x	x	x		x		x								
Kazakh	KK	3	x	x	x		x		x								
Khmer	KM	3	x	x	x		x		x								
Korean	KO	2	x	x	x		x	x	x	x			x				
Kyrgyz	KY	3	x	x	x		x		x								
Laos	LO	3	x	x	x		x		x								
Latvian	LV	2	x	x	x		x		x								
Lithuanian	LT	2	x	x	x		x		x								
Macedonian	MK	2	x	x	x		x		x								
Malay	MS	2	x	x	x		x		x								
Malayalam	ML	3	x	x	x		x		x								
Marathi	MR	3	x	x	x		x		x								
Mongolian	MN	3	x	x	x		x		x								
Nepali	NE	3	x	x	x		x		x								
Norwegian Bokmal	NB	1	x	x	x		x	x	x	x			x				
Norwegian Nynorsk	NN	1	x	x	x		x	x	x	x			x				

Oriya	OR	3	x	x	x		x		x								
Persian / Farsi	FA	3	x	x	x		x		x	x			x				
Polish	PL	2	x	x	x		x		x	x		x					
Portuguese	PT	1	x	x	x		x		x	x	x	x	x	x	x	x	x
Punjabi	PA	3	x	x	x		x		x								
Romanian	RO	2	x	x	x		x		x	x		x					
Russian	RU	2	x	x	x		x		x	x		x	x				
Serbian	SR	2	x	x	x		x		x	x		x					
Sindhi	SD	3	x	x	x		x		x								
Sinhala	SI	3	x	x	x		x		x								
Slovak	SK	2	x	x	x		x		x	x		x					
Slovenian	SL	2	x	x	x		x		x	x		x					
Spanish	ES	1	x	x	x		x		x	x	x	x	x	x	x	x	x
Swahili	SW	3	x	x	x		x		x								
Swedish	SV	1	x	x	x		x	x	x	x		x	x				x
Tagalog / Filipino	TL	3	x	x	x		x		x								
Tamil	TA	3	x	x	x		x		x								
Telugu	TE	3	x	x	x		x		x								
Thai	TH	3	x	x		x	x		x								
Turkish	TR	3	x	x	x		x		x	x							
Ukrainian	UK	2	x	x	x		x		x	x		x					
Urdu	UR	3	x	x	x		x		x								
Uzbek	UZ	3	x	x	x		x		x								
Vietnamese	VI	3	x	x		x	x		x	x							
Welsh	CY	3	x	x	x		x		x								
Zulu	ZU	3	x	x	x		x		x								
TOTAL LANGUAGES	**78**		**78**	**78**	**74**	**4**	**78**	**6**	**78**	**28**	**8**	**21**	**15**	**9**	**7**	**8**	**9**

Linguistic Services Matrix – Language Variants

LANGUAGE VARIANT	ISO	TIER	LANGID	SENTENCE SEG	TOKENIZATION	WORD SEG	LEMMATIZATION	DECOMPOUNDING	SPELLING	POS TAGGING	PHRASE	PARSE	ENTITIES	SENTIMENT	CATEGORIZATION	TOPIC	BOTS & ASSISTANTS
Arabic (MSA)	AR	3	x	x	x		x		x	x			x				
Arabic (Egypt)	AR	3	x	x	x		x		x				x				
Arabic (Gulf)	AR	3	x	x	x		x		x				x				
Arabic (Najdi)	AR	3	x	x	x		x		x				x				
Chinese (Simplified)	ZH	3	x	x		x	x		x	x		x	x	x			
Chinese (Traditional)	ZH	3	x	x		x	x		x	x		x	x	x			
Dutch (Netherlands)	NL	1	x	x	x		x	x	x	x	x	x	x	x	x	x	x
Dutch (Belgium)	NL	1	x	x	x		x	x	x	x	x	x	x	x	x	x	x
English (USA)	EN	1	x	x	x		x		x	x	x	x	x	x	x	x	x
English (UK)	EN	1	x	x	x		x		x	x	x	x	x	x	x	x	x
English (India)	EN	1	x	x	x		x		x	x	x	x	x	x	x	x	x
Finnish (Standard)	FI	3	x	x	x		x		x								
Finnish (Colloquial)	FI	3	x	x	x		x		x								
French (France)	FR	1	x	x	x		x		x	x	x	x	x	x	x	x	x
French (Canada)	FR	1	x	x	x		x		x	x	x	x	x	x	x	x	x
French (Switzerland)	FR	1	x	x	x		x		x	x	x	x	x	x	x	x	x
German (Germany)	DE	1	x	x	x		x	x	x	x	x	x	x	x	x	x	x
German (Switzerland)	DE	1	x	x	x		x	x	x	x	x	x	x	x	x	x	x
Italian (Italy)	IT	1	x	x	x		x		x	x	x	x	x	x	x	x	x
Italian (Switzerland)	IT	1	x	x	x		x		x	x	x	x	x	x	x	x	x
Portuguese (Portugal)	PT	1	x	x	x		x		x	x	x	x	x	x	x	x	x

Portuguese (Brazil)	PT	1	x	x	x		x		x	x	x	x	x	x	x	x	x
Spanish (Spain)	ES	1	x	x	x		x		x	x	x	x	x	x	x	x	x
Spanish (North America)	ES	1	x	x	x		x		x	x	x	x	x	x	x	x	x
Spanish (Central America)	ES	1	x	x	x		x		x	x	x	x	x	x	x	x	x
Spanish (Andes)	ES	1	x	x	x		x		x	x	x	x	x	x	x	x	x
Spanish (Southern Cone)	ES	1	x	x	x		x		x	x	x	x	x	x	x	x	x
TOTAL VARIANTS	**27**		**27**	**27**	**25**	**2**	**27**	**4**	**27**	**22**	**19**	**21**	**25**	**21**	**19**	**10**	**10**
TOTAL LANGUAGES AND VARIANTS	**105**		**105**	**105**	**99**	**6**	**105**	**10**	**105**	**50**	**27**	**42**	**40**	**30**	**26**	**18**	**19**