Bitext Lexical Data Resources

Bitext Lexical Data Resources are the most comprehensive and consistent set of language data resources in the world, with support for 100+ languages and dialects. This proprietary data has been developed to meet the highest quality standards in the field of computational linguistics. Bitext data is used in production by some of the world's largest and most successful software companies.

Use cases

Bitext language data can be leveraged to build high performance text analytics components and functionality across a wide range of software products and development tools. Our data provides companies with the ability to rapidly develop high performance Natural Language Processing (NLP) components such as lemmatizers, POS taggers, phrase extractors, parsers, etc. Bitext data can also be leveraged by software development teams to develop new features and functionality, and add language support for applications that rely on the understanding of text. In particular, applications in the field of Natural Language Understanding (NLU) and Artificial Intelligence (AI) can leverage Bitext data. Some applications that can benefit from Bitext language data include search, mobile keyboards, virtual agents, chatbots, spell checking and grammar checking.

Features

Bitext data sets are rich with comprehensive features. Each language resource has an array of meta data that are relevant to the unique attributes of each specific language and the data features are consistent across all languages. This comprehensiveness and richness in data provides unlimited flexibility, adaptability and customizability. Inflectional morphology, derivational morphology, use variants and word formation are just a few of the features that are covered by the data. The inflected words in each data set are provided with applicable meta tags and/or information such as:

- Lemma: The canonical form for the inflected word is provided.
- POS: Part of Speech such as noun, verb, adjective, etc. is defined.

- Voice: Verb form is classified as active or passive.
- Tense: Specifies when the action takes place such as past, present, future, etc.
- Aspect: Indicates whether the action is complete, ongoing, habitual, etc.
- Mood: Modality of the verb form is provided: indicative, subjunctive, imperative, etc.
- Person: Verb or pronoun refers to the first, second or third person.
- Number: State of being singular, dual or plural.
- Gender: Noun, verb or adjective forms are provided, masculine, feminine, neuter, etc.
- Case: The function that the noun or adjective plays within a sentence.
- Degree: An adjective is specified as in its positive, comparative or superlative form.
- Definiteness: Specifies whether a noun or adjective refers to a concrete or general concept.
- Polarity: Indicates whether a verb, adjective or noun is in a negative form.
- Contractions: Shortened form of a word or group of words are provided.
- Pronominal Clitics: Clitic pronouns are identified and tagged.
- Formality: Indicates the social status of the speaker in relation to the context.
- Frequency: Relative frequency of the form based on a large general-purpose corpus.
- Named Entities: Pre-defined entities are tagged as person names, places, organization, etc.
- Offensive: Indicates whether the form might be considered offensive in certain contexts.

LXD Feature Matrix

LANGUAGE	ISO	TIER	LEMMA	POS	VOICE	TENSE	ASPECT	MOOD	PERSON	NUMBER	GENDER	CASE	DEGREE	DEFINITENESS / STATE	NEGATIVE	CONTRACTIONS	PRONOMINAL CLITICS	FORMALITY	FREQUENCY	NAMED ENTITIES	OFFENSIVE	CATEGORY	TOTAL NUMBER OF LEMMAS	TOTAL NUMBER OF FORMS
Afrikaans	AF	1	x	x		x	x		x	x	x	x	x			x			x	x	x		20 K	38 K
Albanian	SQ	2	x	x	x	x		x	x	x	x	x		x					x	x	x		35 K	284 K
Amharic	AM	3	x	x		x			x	x	x	x		x	x			x	x	x	x		16 K	230 K
Arabic	AR	3	x	x	x	x		x	x	x	x	x		x			x		x	x	x		22 K	17 M
Armenian	HY	2	x	x		x		x	x	x		x	x	x			x		x	x	x		6 K	150 K
Assamese	AS	2	x	x		x			x	x	x	x			x			x	x	x	x		30 K	1.26 M
Azeri	AZ	3	x	x	x	x		x	x	x		x			x				x	x	x		14 K	1.1 M
Basque	EU	3	x	x		x		x		x		x							x	x	x		45 K	25 M
Belarusian	BE	2	x	x		x	x		x	x	x	x	x						x	x	x		66 K	1 M
Bengali	BN	2	x	x		x		x	x	x		x		x	x			x	x	x	x		54 K	1.47 M
Bulgarian	BG	2	x	x		x			x	x	x	x		x					x	x	x		75 K	800 K
Burmese	MY	3	x	x															x	x	x		30 K	30 K
Catalan	CA	1	x	x		x		x	x	x	x					x	x		x	x	x		35 K	1.5 M
Chinese	ZH	3	x	x															x	x	x		75 K	75 K
Croatian	HR	2	x	x	x	x	x		x	x	x	x	x	x		x			x	x	x		44 K	434 K
Czech	CS	2	x	x	x	x		x	x	x	x	x			x			x	x	x	x		55 K	4 M
Danish	DA	1	x	x	x	x			x	x	x	x	x	x					x	x	x		60 K	700 K
Dutch	NL	1	x	x		x		x	x	x	x					x			x	x	x		90 K	500 K
English	EN	1	x	x		x			x	x	x		x			x			x	x	x		60 K	180 K
Esperanto	EO	1	x	x		x				x		x				x			x	x	x		50 K	400 K

Estonian	ET	3	x	x	x	x		x	x	x		x	x		x				x	x	x		85 K	7 M
Finnish	FI	3	x	x	x	x		x	x	x		x	x				x	x	x	x	x		70 K	80 M
French	FR	1	x	x		x		x	x	x	x					x	x		x	x	x		60 K	1.4 M
Galician	GL	1	x	x		x		x	x	x	x		x				x		x	x	x		45 K	5 M
Georgian	KA	3	x	x		x		x	x	x		x							x	x	x		23 K	500 K
German	DE	1	x	x		x		x	x	x	x	x	x			x			x	x	x		100 K	2.5 M
Greek	EL	2	x	x	x	x	x	x	x	x	x	x	x					x	x	x	x		27 K	500 K
Gujarati	GU	3	x	x		x	x		x	x	x	x							x	x	x		45 K	2.5 M
Hebrew	HE	3	x	x		x			x	x	x			x			x		x	x	x		23 K	12 M
Hindi	HI	2	x	x		x		x	x	x	x	x					x		x	x	x		30 K	500 K
Hungarian	HU	3	x	x		x		x	x	x		x	x				x		x	x	x		75 K	18 M
Icelandic	IS	2	x	x	x	x		x	x	x	x	x	x						x	x	x		50 K	1.75 M
Indonesian	ID	1	x	x	x		x			x			x				x		x	x	x		35 K	150 K
Irish Gaelic	GA	2	x	x		x		x	x	x	x	x	x			x			x	x	x	x	30 K	1.5 M
Italian	IT	1	x	x		x		x	x	x	x					x	x		x	x	x		65 K	1.4 M
Japanese	JP	3	x	x	x	x												x	x	x	x		450 K	9.4 M
Kannada	KN	3	x	x		x		x	x	x	x	x			x				x	x	x		40 K	500 K
Kazakh	KK	3	x	x	x	x	x	x	x	x		x	x		x		x		x	x	x		10 K	2 M
Khmer	KM	3	x	x															x	x	x		30 K	30 K
Korean	KO	2	x	x	x	x		x				x						x	x	x	x		75 K	6.25 M
Kyrgyz	KY	3	x	x	x	x	x	x	x	x		x	x		x		x		x	x	x		10 K	2 M
Laos	LO	3	x	x		x													x	x	x		45 K	45 K
Latvian	LV	2	x	x	x	x	x	x	x	x	x	x	x	x	x				x	x	x		42 K	2.37 M
Lithuanian	LT	2	x	x		x		x	x	x	x	x	x	x	x				x	x	x		44 K	26 M
Macedonian	MK	2	x	x		x	x		x	x	x		x	x					x	x	x		30 K	150 K
Malay	MS	1	x	x	x		x			x			x				x		x	x	x		45 K	120 K
Malayalam	ML	3	x	x		x		x	x	x	x	x			x			x	x	x	x		35 K	500 K
Marathi	MR	2	x	x	x	x	x	x		x	x	x						x	x	x	x		19 K	17 M
Mongolian	MN	3	x	x		x	x	x	x	x		x	x		x				x	x	x		23 K	500 K
Nepali	NE	3	x	x	x	x			x	x	x	x	x		x		x	x	x	x	x		15 K	1 M

Norwegian Bokmal	NB	1	x	x		x			x	x	x	x	x	x					x	x	x		45 K	500 K
Norwegian Nynorsk	NN	1	x	x		x			x	x	x	x	x	x					x	x	x		75 K	400 K
Oriya	OR	2	x	x		x	x		x	x		x		x				x	x	x	x		33 K	63 K
Persian	FA	3	x	x		x	x	x	x	x			x	x	x		x		x	x	x		10 K	400 K
Polish	PL	2	x	x	x	x	x	x	x	x	x	x							x	x	x		95 K	1.45 M
Portuguese	PT	1	x	x		x		x	x	x	x						x		x	x	x		40 K	3.5 M
Punjabi	PA	3	x	x		x	x		x	x	x	x							x	x	x		20 K	240 K
Romanian	RO	2	x	x	x	x		x	x	x	x	x		x					x	x	x		36 K	300 K
Russian	RU	2	x	x		x		x	x	x	x	x	x						x	x	x		50 K	1.5 M
Serbian	SR	2	x	x	x	x	x		x	x	x	x	x	x		x			x	x	x		45 K	1.5 M
Sindhi	SD	2	x	x	x	x		x	x	x	x	x						x	x	x	x		17 K	451 K
Sinhala	SI	2	x	x		x		x	x		x	x		x				x	x	x	x		30 K	916 K
Slovak	SK	2	x	x		x	x		x	x	x	x	x		x				x	x	x		45 K	1.5 M
Slovenian	SL	2	x	x		x	x	x	x	x	x	x	x	x					x	x	x		22 K	178 K
Spanish	ES	1	x	x		x		x	x	x	x						x		x	x	x		60 K	2.5 M
Swahili	SW	3	x	x		x			x	x									x	x	x		34 K	650 K
Swedish	SV	1	x	x	x	x		x	x	x	x	x	x	x					x	x	x		70 K	500 K
Tagalog	TL	2	x	x	x		x	x											x	x	x		40 K	90 K
Tamil	TA	2	x	x		x			x	x	x	x					x		x	x	x		27 K	1 M
Telugu	TE	3	x	x		x	x		x	x	x	x			x				x	x	x		30 K	1.5 M
Thai	TH	2	x	x														x	x	x	x		40 K	40 K
Turkish	TR	3	x	x		x		x	x	x		x			x		x		x	x	x		300 K	3.5 M
Ukrainian	UK	2	x	x		x	x	x	x	x	x	x	x				x		x	x	x		40 K	650 K
Urdu	UR	2	x	x		x		x	x	x	x	x					x		x	x	x		15 K	200 K
Uzbek	UZ	3	x	x	x	x	x	x	x	x		x	x		x		x		x	x	x		11 K	1 M
Vietnamese	VI	2	x	x															x	x	x		34 K	40 K
Zulu	ZU	3	x	x		x		x		x	x	x							x	x	x		10 K	1 M
TOTAL LANGUAGES	**77**		**77**	**77**	**26**	**69**	**25**	**46**	**62**	**68**	**49**	**55**	**33**	**21**	**19**	**11**	**23**	**15**	**77**	**77**	**77**	**1**	**77**	**77**

LXD Feature Matrix

LANGUAGE VARIANT	ISO	TIER -LEXICAL	LEMMA	POS	VOICE	TENSE	ASPECT	MOOD	PERSON	NUMBER	GENDER	CASE	DEGREE	DEFINITENESS/ STATE	NEGATIVE	CONTRACTIONS	PRONOMINAL CLITICS	FORMALITY	FREQUENCY	NAMED ENTITIES	OFFENSIVE	CATEGORY	TOTAL NUMBER OF LEMMAS	TOTAL NUMBER OF FORMS
Arabic (MSA)	AR	3	x	x	x	x		x	x	x	x	x		x			x		x	x	x		22 K	17.8 M
Arabic (Gulf)	AR	3	x	x	x	x		x	x	x	x	x		x			x		x	x	x		22 K	9 M
Arabic (Najdi)	AR	3	x	x	x	x		x	x	x	x	x		x			x		x	x	x	x	23 K	1 M
Chinese (Simplified)	ZH	3	x	x															x	x	x		74 K	74 K
Chinese (Traditional)	ZH	3	x	x															x	x	x		74 K	74 K
Dutch (Netherlands)	NL	1	x	x		x		x	x	x	x					x			x	x	x		106 K	586 K
Dutch (Belgium)	NL	1	x	x		x		x	x	x	x					x			x	x	x		97 K	591 K
English (USA)	EN	1	x	x		x			x	x	x		x			x			x	x	x		63 K	188 K
English (UK)	EN	1	x	x		x			x	x	x		x			x			x	x	x		63 K	190 K
English (India)	EN	1	x	x		x			x	x	x		x			x			x	x	x		65 K	193 K
Finnish (Standard)	FI	3	x	x	x	x		x	x	x		x	x				x	x	x	x	x		74 K	74 M
Finnish (Colloquial)	FI	3	x	x	x	x		x	x	x		x	x				x	x	x	x	x		71 K	22.6 M
French (France)	FR	1	x	x		x		x	x	x	x					x	x		x	x	x		76 K	1.45 M
French (Canada)	FR	1	x	x		x		x	x	x	x					x	x		x	x	x		62 K	1.47 M
French (Switzerland)	FR	1	x	x		x		x	x	x	x					x	x		x	x	x		62 K	1.47 M
German (Germany)	DE	1	x	x		x		x	x	x	x	x	x			x			x	x	x		101 K	2.6 M
German (Switzerland)	DE	1	x	x		x		x	x	x	x	x	x			x			x	x	x		108 K	2.5 M
Italian (Italy)	IT	1	x	x		x		x	x	x	x					x	x		x	x	x		82 K	1.47 M
Italian (Switzerland)	IT	1	x	x		x		x	x	x	x					x	x		x	x	x		70 K	1.48 M
Portuguese (Portugal)	PT	1	x	x		x		x	x	x	x						x		x	x	x		51 K	3.78 M
Portuguese (Brazil)	PT	1	x	x		x		x	x	x	x						x		x	x	x		36 K	3 M

Spanish (Spain)	ES	1	x	x		x		x	x	x	x						x		x	x	x		85 K	1.34 M
Spanish (North America)	ES	1	x	x		x		x	x	x	x						x		x	x	x		58 K	1.25 M
Spanish (Central America)	ES	1	x	x		x		x	x	x	x						x		x	x	x		58 K	1.24 M
Spanish (Andes)	ES	1	x	x		x		x	x	x	x						x		x	x	x		59 K	1.25 M
Spanish (Southern Cone)	ES	1	x	x		x		x	x	x	x						x		x	x	x		59 K	1.32 M
TOTAL VARIANTS	**25**		**26**	**25**	**5**	**24**	**-**	**21**	**24**	**24**	**22**	**7**	**4**	**3**	**-**	**12**	**17**	**2**	**26**	**26**	**26**		**26**	**26**
TOTAL LANGUAGES AND VARIANTS	**103**		**103**	**103**	**32**	**93**	**26**	**67**	**86**	**92**	**71**	**73**	**37**	**24**	**19**	**23**	**40**	**18**	**103**	**103**	**103**	**1**	**103**	**103**

Key to Feature Matrix

voice	indicates whether the verb form is active or passive
tense	indicates when the action happens (past, present, future, ...)
aspect	indicates whether the action is complete, ongoing, habitual, …
mood	indicates the modality of the verb form (indicative, subjunctive, imperative, …)
person	indicates whether the verb form or pronouns refers to the first (speaker), second (listener) or a third person
number	indicates whether the form is singular, dual, plural, …
gender	indicates whether the form is masculine, feminine, neuter, …
case	indicates the relationship of the noun/adjective to other words in the sentence
degree	indicates whether an adjective is in positive, comparative or superlative form
definiteness/state	indicates whether a noun/adjective is in indefinite, definite or construct form
negative	indicates whether a verb/adjective is in positive or negative form
contractions	common contractions, including negation, articles, …
pronominal clitics	a clitic pronoun, often used to mark objects (for verbs) or possessives (for nouns/adjectives)
declension type	indicates whether the noun/adjective is short, long, …
formality	indicates the relative social status of the speaker and, optionally, the listener or a third person
frequency	indicates the relative frequency of the form in a large, general-purpose corpus
named entities	person names, places, companies, organizations, …
offensive	indicates whether the form might be considered offensive in certain contexts