Kalmasoft KTagset Technical Documentation

Draft revision 1.1, July 19 2025

Author: Kalmasoft

== Work in progress, may contain errors ==

Table of contents

1. About this document

1.1 Purpose of this document

This document defines a custom Part of Speech (POS) tagset designed for NLP applications; the suggested tagset balances cross-linguality, granularity and compactness.

1.2 Scope of this document

This document is prepared for linguists, system developers, and IT specialists.

1.3 Legal notice

1.4 Content note

Contents about languages, definitions, and linguistic information in this document generated using Google Gemini, while careful revision is made to check every part generated some information may still be irrelevant.

1.5 Disclaimer of endorsement

2. Part-of-Speech tagset

2.1 Core concept

POS tagging is the process of assigning grammatical categories to individual words in a text corpus. Each word receives a tag that identifies its part of speech based on both its definition and context within the sentence.

2.2 Overview

Part-of-Speech (POS) tagsets are standardized predefined collections of labels used to categorize words based on their grammatical roles and syntactic functions within sentences. These tagsets serve as the foundation for automated tasks like text analysis (syntactic, semantic), natural language processing applications including linguistic research, machine learning, machine translation, sentiment analysis, cross-lingual information retrieval, corpus automatic annotation, and information extraction.

Kalmasoft's Tagset (KTagset©) is a cross-lingual generic tagset, a collection of tags provide a standardized way to represent grammatical roles of words and sentences in a corpus by describing the semantic, syntactic and morphological features of any predictable surface form, it is the default tagset used by Kalmasoft's in products including (Arabic *PoS Tagger, Arabic Verb Conjugator, Arabic Nouns Inflector*, and *Arabic Named Entity Recognition*), KTagset is compact and machine-oriented with adequate human readability.

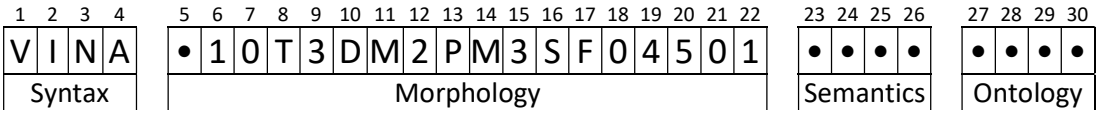
KTagset is not a “universal PoS tagset”, not meant to substitute any custom part of speech tagset currently in use; it is best described as a “set of uniform pos labeling tags”.

3. Tagging scheme

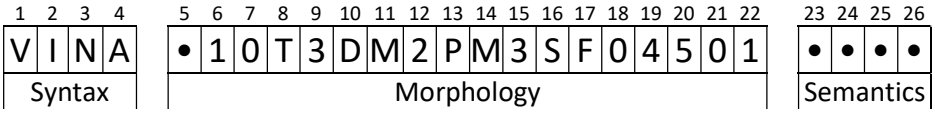
3.1 KTag design

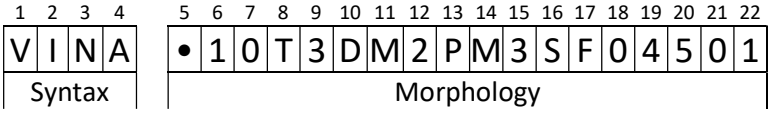
A KTag is string of thirty [30] characters grouped in three 4-characters sets and one 18-characters set [4-18-4-4], the 18 characters set ends with five place holders reserved to represent optional language variety or dialect while the rest of characters have their different grammatical notation, each letter may have different representation depending on the set it resides in and the specific position it occupies, the tag string is basically case sensitive with optional lower case or Greek letters for different levels of detail and extended grammatical features, broadly split into {syntax}{morphology}{semantics}{ontology}.

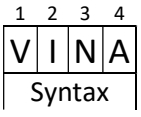
KTag Canonical form: VINA•10T3DM2PM3SF04501•••••••



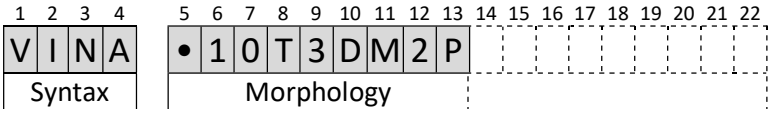
Short forms, break only at group boundaries.







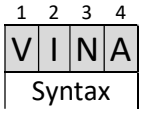
Custom form, variable length minimum required tags.



3.2 KTag listing

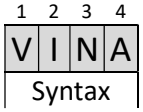
Each tag in the set is defined in the (tagging scheme) explaining the words and contexts where it applies, both “Short forms” and “Custom forms” are used throughout this document for examples and use cases, the contents inside the tables are ordered by most common use first.

3.2.1 Major class tags



Cell	Tag	Category	Description	Examples
1	V	Verbs	Verb	run, solve, create
1	D	Adverbs	Adverbs	quickly, very, efficiently
1	N	Nouns	Noun	book, water, happiness
1	R	Pronouns	Pronouns	it, they, this
1	J	Adjectives	Adjectives	
1	P	Particles	Particle	to, not
1	F	Affixes	Affixes	
1	A	Determiners	Articles	the
1	C	Conjunctions	Conjunctions	and, if
1	I	Interjections	Interjections	
1	E	Entities	Entity	
1	S	Symbols	Mathematical, currency etc	$, %, +
1	M	Numbers	Number, numerals	5, hundred, third
1	L	Scripts	Language script, writing scripts	
1	T	Technical	Technical	VPN, TCP IP

3.2.2 Syntactic feature tags



Cell	Tag	Category	Description	Examples
2	P	Tenses	Aorist, preterite, past	
2	I	Tenses	Imperfect, present	
2	F	Tenses	Future	
			Progressive?	
2	P	Aspects	Perfective	
2	I	Aspects	Imperfective	
2	A	Nouns	Acronyms, abbreviations	Sarl, GmbH, LLC, Ltd.
2	M	Nouns	Verbal noun, Arabic Masdar	
2	C	Nouns	Active participle	
2	P	Nouns	Passive participle	
2	H	Nouns	Hyperbolic participle	
2	S	Nouns	Superlative	
2	I	Nouns	Noun of instrument, Arabic?	
2	B	Nouns	Broken plural, Arabic	
2	P	Particles	Preposition	
2	A	Particles	Adposition	in, on, under
2	R	Particles	Interrogative	
2	N	Particles	Conditional	
2	M	Particles	Simile	
2	E	Particles	Emphasis	
2	X	Particles	Exclusion	
2	S	Particles	Postposition	
2	P	Affixes	Prefix	
2	S	Affixes	Suffix	
2	I	Affixes	Infix	
2	D	Determiners	Definite article	
2	I	Determiners	Indefinite article	
2	R	Determiners	Proper article	
2	P	Determiners	Partitive article	
2	N	Determiners	Negative article	
2	C	Conjunctions	Coordinating, joins words/clauses	and, or, but
2	S	Conjunctions	Subordinating	if, because, while
2	E?	Interjections	Exclamation	oh, wow, hello
2	P	Symbols	Punctuation, sentence separators	., ,, ?, ;
2	C	Symbols	Currency	
2	M	Symbols	Mathematical	
2	D	Symbols	Date	
2	T	Symbols	Time	
3	I	Moods	Indicative	

Cell	Tag	Category	Description	Examples
3	S	Moods	Subjunctive	
3	M	Moods	Imperative	
3	J	Moods	Jussive	
3	E	Moods	Emphatic	
3	R	Moods	Interrogative	
3	N	Cases	Nominative	
3	A	Cases	Accusative	
3	G	Cases	Genitive, possessive	
3	D	Cases	Dative	
3	I	Cases	Instrumental	
3	L	Cases	Locative	
3	P	Cases	Prepositional	
3	C	Cases	Causative	
3		Cases	Comparative	
3		Cases	Equative	
3		Cases	Superlative	
3	R	Cases	Partitive, Finnish	
3		Cases	Inessive, Finnish, Hungarian	
3	E	Cases	Elative, Finnish, Hungarian	“out of”
3		Cases	Illative, Finnish, Hungarian	“into”
3		Cases	Adessive, Finnish, Hungarian	“at”
3	O	Cases	Ablative, Finnish, Hungarian	“away from”
3	U	Cases	Instructive, Finnish	
3	S	Cases	Essive, Finnish	
3	T	Cases	Translative, Hungarian	
3	C	Cases	Comitative, Finnish	
3	B	Cases	Abessive, Finnish	
3		Cases	Allative, Hungarian	
3		Cases	Sublative, Hungarian	“onto”
3		Cases	Delative, Hungarian	“from off, about”
3		Cases	Distributive, Hungarian	
3		Cases	Sociative, Hungarian	
3		Cases	Suppressive, Hungarian	
3	X	Cases	Terminative, Hungarian	
3	D	Affixes	Derivational	de-, ex-, -y
3	I	Affixes	Inflectional	-s,
4	A	Voices	Active voice	
4	P	Voices	Passive voice	
4	C	Voices	Causative	
4	R	Voices	Reflexive, Greek	
4		Voices	Causative passive, Hebrew	
4	I	Voices	Intensive, Hebrew	
4		Voices	Intensive passive, Hebrew	
4	T	Adverbs	Temporal adverb	
4	L	Adverbs	Locative adverb	

Cell	Tag	Category	Description	Examples
4	C	Adverbs	Comparative adverb	
4	S	Adverbs	Superlative adverb	
4	E	Nouns?	Relative nominal construct, Arabic?	
4	D	Nouns	Diminutive	
4	N	Nouns	Proper noun	
4	Q	Nouns?	Numerative noun, Arabic?	
4	H	Nouns	Instance noun, Arabic?	
4	T	Nouns	Temporal noun, Arabic?	
4	L	Nouns	Locative noun, Arabic	Butchery, laundry
4	C	Nouns?	Cardinal number	
4	O	Nouns?	Ordinal number	
4	U	Pronouns	Personal pronoun	
4	W	Pronouns	Relative pronoun	
4	M	Pronouns	Demonstrative pronoun	
4	X	Pronouns	Reflexive pronoun	
4	S	Pronouns	Possessive pronoun	
4	C	Pronouns	Reciprocal pronoun	
4	R	Adjectives	Relative adjective, Arabic?	
4	Q	Adjectives	Quasiverbal adjective, Arabic?	
4	R		Interrogative	
			Noun classes, Bantu languages	
4	N	Affixes	Noun affix	
4	L	Affixes	Locative affix	
4	O?	Affixes	Negative affix	
4		Affixes	Neuter affix	
4	R	Affixes	Pronoun affix	
4		Affixes	Progressive affix	
4	X	Affixes	Reflexive affix	
4		Affixes	Relative affix	
4	G	Affixes	Interrogative affix	
4	M	Affixes	Diminutive affix	
4		Affixes	Augmentative affix	
4		Affixes	Copulative affix	
4	D	Affixes	Adverbial affix	
4	U	Affixes	Personal affix	

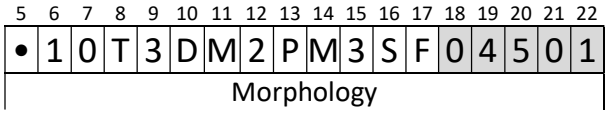
3.2.3 Morphological feature tags

5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
•	1	0	T	3	D	M	2	P	M	3	S	F	0	4	5	0	1
Morphology																	

Cell	Tag	Category	Description	Examples
5			Participles, progressive?	
5	F	Nouns	Definite	
5	I	Nouns	Indefinite	
6-7		Forms	Surface forms, Bantu noun classes	
8	I	Valency	Monovalent, Intransitive verb	
8	T	Valency	Divalent, Transitive verb	
8	D	Valency	Trivalent, Ditransitive verb	
8	A	Valency	Ambitransitive verb	
8		Valency	Reflexive verb	
8	R	Valency	Reciprocal verb	
8	C	Valency	Causative verb	
8		Valency	Applicative verb	
8	X	Valency	Auxiliary verb?	
8				
9	1	Person	1st person subject, possessor	
9	2	Person	2nd person, subject, possessor	
9	3	Person	3rd person, subject, possessor	
12	1	Person	1st person, first object, possessive	
12	2	Person	2nd person, first object, possessive	
12	3	Person	3rd person, first object, possessive	
15	1	Person	1st person, second object	
15	2	Person	2nd person, second object	
15	3	Person	3rd person, second object	
10	S	Number	Singular, subject, possessor	
10	D	Number	Dual, subject, possessor	
10	P	Number	Plural, subject, possessor	
13	S	Number	Singular, first object, possessive	
13	D	Number	Dual, first object, possessive	
13	P	Number	Plural, first object, possessive	
16	S	Number	Singular, second object	
16	D	Number	Dual, second object	
16	P	Number	Plural, second object	
11	M	Gender	Masculine, subject, possessor	
11	F	Gender	Feminine, subject, possessor	
11	U	Gender	Unisex, subject, possessor	
11	G	Gender	GLPT, subject, possessor	

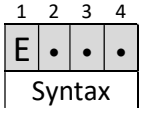
Cell	Tag	Category	Description	Examples
11	N	Gender	Neuter, subject, possessor	
14	M	Gender	Masculine, first object, possessive	
14	F	Gender	Feminine, first object, possessive	
14	U	Gender	Unisex, first object, possessive	
14	G	Gender	GLPT, first object, possessive	
14	N	Gender	Neuter, first object, possessive	
17	M	Gender	Masculine, second object	
17	F	Gender	Feminine, second object	
17	U	Gender	Unisex, second object	
17	G	Gender	GLPT, second object	
17	N	Gender	Neuter, second object	

3.2.4 Morphological feature, language tags



Cells	Tag	Category	Description	Examples
18-20		Languages	ISO 639-3 Alpha-3 code*	amh:Amharic, eng:English
18-22		Varieties	HIS ROLV codes*	06442:Levantine Arabic
18-22		Varieties	IETF BCP 47 language tag	
21-22		Countries	ISO 3166-1 alpha-2, entities	US: United States
21-22	F•	Politeness	Formal	
21-22	I•	Politeness	Informal	
21-22	LM?	Politeness	Normal	
21-22	H•	Politeness	Honorific	
21-22	G•	Politeness	Derogatory	
21-22	D•	Politeness	Direct	
21-22	R•	Politeness	Indirect	
21-22	V•	Politeness	Vulgar	
21-22	X•	Politeness	Obscene	
21-22	O•	Politeness	Profane	
21-22	S•	Politeness	Swear word	
21-22	P•	Politeness	Positive	
21-22	N•	Politeness	Negative	

3.2.5 Named Entity tags



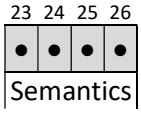
Entities come in three different levels E1, E2 and E3 with E1 being sufficient for basic NER applications; E2 and E3 are optional subsets ready to implement and fit any other sophisticated project.

Cells	Tag	Category	Description	Examples
2-4	PER		E1 Person	
2-4	LOC		E1 Location	
2-4	ORG		E1 Organization	
2-4	GRP		E1 Group	Muslims, Haitian
2-4	PRO		E1 Product	
2-4	VNT		E1 Event, state	
2-4	QTY		E1 Quantity	
2-4	SUB		E1 Substance	
2-4	ANM		E1 Animate	
2-4	STR	?	E2 Astronomy	Polaris, Pluto, Moon
2-4	REC	?	E3 Recurring	
2-4	HMN	Animate	E2 Human	
2-4	NML	Animate	E2 Animal	
2-4	PLT	Animate	E2 Plant?	
2-4	ACT	Events	E2 Activity	
2-4	NAT	Events	E2 Natural	
2-4	HMD	Events	E2 Human-made	
2-4	DIS	Events	E2 Disease	
2-4	EMO	Events	E2 Emotion	
2-4	STN	Events	E2 Situation	
2-4	GEO	Groups	E2 Geography related	
2-4	NTL	Groups	E2 Nationality	
2-4	PLC	Groups	E2 Political	
2-4	RLG	Groups	E2 Religious?	
2-4	ETH	Groups	E2 Ethnicity	
2-4	FAC	Locations	E2 Facility	
2-4	PPL	Locations	E2 Populated place	
2-4	MRK	Locations	E2 Landmark	
2-4	MAS	Locations	E2 Landmass	
2-4	CLS	Locations	E2 Celestial	
2-4	COR	Organizations	E3 Corporation	
2-4	TTL	Personal	E2 Title	
2-4	FAM	Personal	E3 Famous	

Cells	Tag	Category	Description	Examples
2-4	FIC	Personal	E2 Fictional	
2-4	JOB	Personal	E3 Professional title	
2-4	RLG	Personal	E3 Religious title	
2-4	HON	Personal	E3 Honorific title	
2-4	CHM	Products	E3 Chemical	
2-4	TAX	Products?	E3 Taxonomy	
2-4	FOD	Products	E3 Food	
2-4	WPN	Products	E3 Weapon	
2-4	VHC	Products	E3 Vehicle	
2-4	LAN	Products	E3 Language	
2-4	CUR	Products	E3 Currency	
2-4	IPT	Products	E2 Intellectual property	
2-4	BOK	Products	E3 Book	
2-4	CON	Products	E3 Concept	
2-4	IDA	Products	E3 Idea	
2-4	PLY	Products	E3 Play	
2-4	PNT	Products	E3 Painting	
2-4	SNG	Products	E3 Song	
2-4	SFT	Products	E3 Software	
2-4	THR	Products	E3 Theory	
2-4	TMK	Products	E3 Trademark	
2-4	MAT	Products	E2 Material	
2-4	CHR	Products	E3 Character	
2-4	PRX	Quantities	E2 Approximation	
2-4	DAT	Quantities	E2 Date	
2-4	TIM	Quantities	E2 Time	
2-4	NUM	Quantities	E2 Number	
2-4	PRC	Quantities	E2 Percent	
2-4	ORD	Quantities	E2 Ordinal	
2-4	CRD	Quantities	E2 Cardinal	
2-4	MON	Quantities	E3 Money	
2-4	DIG	Quantities	E3 Digit	
2-4	DUR	Quantities	E3 Duration	
2-4	WGT	Quantities	E3 Weight	
2-4	RNG	Quantities	E3 Range	
2-4	AGE	Quantities	E3 Age	
2-4	REA	Quantities	E3 Area	
2-4	VOL	Quantities	E3 Volume	
2-4	GAS	Substances	E2 Gas	
2-4	LQD	Substances	E2 Liquid	
2-4	SLD	Substances	E2 Solid	

3.2.6 Semantic feature tags

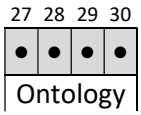
Semantic features deal with concepts like negativity and pragmatics.



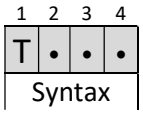
Cell	Tag	Category	Description	Examples
23		Negation		
24	R		Interrogative	
24	D		Declarative	
24	X		Exclamatory	
24	M		Imperative	
25				
26				

3.2.7 Ontology feature tags

This section is blank.



3.2.8 Special technical tags



Cells	Tag	Category	Description	Examples
2	C	Codes	Code snippets	`print()`, `var_name`, `SELECT`
2	T	Terms	Technical terms, acronyms	`API`, `JSON`, `blockchain`
2	W		Web addresses	`https://example.com`
2	S		Emails, hashtags, social media	`@admin`, `@user_123`, `#kalmasoft`, `#wego`

3.2.9 Special common tags

These tags are common to all cells, not applicable bullet symbol is context-sensitive.

Tag	Class	Description	Examples
#	Wildcard	Not applicable, place holder	
*****	Wildcard	Collective	
?	Unparsed	Unknown words, foreign terms, typo	`asdf`, `lol`, `HTTP/2`
•	Holder	Not applicable, bullet, Unicode U+2022	
∅	Holder	Nothing, U+2205	

4. Integrating representational standards

Optional representation standards may be integrated with KTagset, documentation of the integrated standard is the sole responsibility of the user.

4.1 Integration of country and language variety standards

Standard	KTag cells	Example	
BPC-47	[18-22]	ja-Kana-JP, ru-Cyrl-BY	Language verities
IETF	[18-22]	sr-Cyrl, zh-Hant-TW	Language verities
ISO 3166-1 alpha-2	[18-22]	SD, DZ, US	Country names
ISO 3166-1 alpha-3	[18-22]	SDN, DZA, USA	Country names
ISO 639-3	[18-22]	amh, nep, eng	Language names
ROLV	[18-22]	03558	Language verities

5. Language-specific considerations

KTagset developed with cross-linguality in mind, suited to this role by design.

5.1 Morphologically rich languages

Languages with complex morphology, also known as morphologically rich languages, are those where grammatical relations are primarily indicated by changes to the words themselves, rather than through word order or function words like prepositions or articles, they often feature intricate systems of inflection, derivation, and compounding. These languages often feature a high degree of inflection, where words can have many different forms to express grammatical categories like tense, number, gender, case, and person. Examples of languages with complex inflectional systems require extended tagsets:

German: Incorporates case, gender, and number information

Turkish: Includes extensive agglutination patterns

Russian: Handles complex case and aspect systems

5.2 Languages with extensive grammatical categories

Languages with large case systems include Hungarian with up to 34 cases, depending on the analysis, Finnish (15), and Estonian (14), cases are formed by adding suffixes to the noun stem.

6. Application examples

6.1 Syntactic analysis examples

Syntactic analysis, also known as parsing, is a crucial step in Natural Language Processing (NLP) that focuses on analyzing the grammatical structure of sentences to understand how words relate to each other and form phrases and clauses.

6.1.1 Basic syntactic usage examples

These examples show basic usage of Custom form KTags without the language variety encoded.

Language	Example	KTag	Gloss
Amharic	**ትናገራለህ**	VINA•01T2SM•••	*you talk*; verb imperfect indicative active
	እንመሰክራለን	VINA•01T1PF•••	*we witness*; verb imperfect indicative active
	ስማኝ	VMNA•01T2SM1SM•••	*listen to me*; verb past imperative active
	ተሰጥቶሃል	VPNP•01T•••2••3S•	*it was given to you*; verb past indicative passive
Arabic	القَلَمُ	N•N••01D•SM••••••	*the pen*; noun nominative singular masculine
	لِمَعْلُومَاتِكَ	N•G••01I2SM3PF•••	*for your information*; noun genitive…
	اسْتِعْرَاضَات	N••••01I•PF••••••	*demonstrations*; noun plural feminine
	يَحْجِزُونَ	VINA•01T3PM••••••	*they reserve*; verb imperfect indicative…
Finnish			
German	unfreundlichkeit	N•N••••••S•••••••	*unfriendliness*;
Hindi			
Japanese	書かない	VINA•••T•S•••••••	*don't write*; verb negative, context free?
Russian	прочитал	VPNA•••T1S•••••••?	*I read*; verb past perfective
	красивого	J•G••••••SM••••••?	*beautiful*; adjective masculine genitive
Swahili	wataamka	VINA•••I3P•••••••	*they will wake up*;
	ameenda	VPNA•••I3SM••••••	*he has gone*;
Turkish	gidiyorum	VINA•••I1S•••••••	*I am going*; verb present continuous tense
	gözlükçü	N•N••••••S•••••••	*optician*; noun
Zulu			

6.1.2 Advanced syntactic usage examples

Examples below show advanced usage of Short form KTags with the language variety ending each tag.

Language	Example	KTag	Gloss
Amharic	**አታስፈራራኝ**	VINA•??T2S•1S••••30535	*don't scare me*;
Arabic	يَعْطِيَانِكُمُوها	VINA•10T3DM2PM3SF04501	*they "both" give it to "all" of you*
Finnish			
German	arbeitslosigkeit		*unemployment*;
Hindi			
Japanese	書きませんでした	VPNA•10T•••••••••jpnH•	*haven't written*; polite past negative
Russian			
Swahili	ninapenda	VINA•01T1SM••••••	*I love*;
Turkish	göndermemişlerdi	VPNA••T3P•••••••?	*they didn't send*;
Zulu	ngiyamkhumbula	VINA•••T1S•2S••••zul••	*I miss you*; I remember you
	bekufika	VP•A•15I3S•••••••?	*it was coming*;

6.1.3 Basic entity usage examples

Examples below show basic usage of Short form KTags with different levels of named entities.

Language	Example	KTag	Gloss
Arabic	ميرسيدس	ETMK	Trademark, Mercedes
English	Amoxillin	EPRO	Product, medicinal
English	Aretha Franklin	EFAM	Famous person
Japanese	大手町駅	EFAC	Facility, Ōtemachi-eki subway station
French	Ibrahim Traoré	EFAM	Famous person
Hebrew	תורת היחסות הפרטית	ETHR	Theory, Special Relativity
Korean	김 정 은	EFAM	Famous person, Kim Jong Un
Syriac	ܓܡܠܐ	ENML	Animal, a camel
Tamazight	ⵜⵉⵣⵉ ⵓⵣⵓ	EPPL	City, Tizi Ouzou, Algeria
Tibetan	सगरमाथा	EMRK	Landmark, Mount Everest

6.1.4 Advanced entity usage examples

Examples below show advanced usage of Short form KTags with different levels of named entities.

Example	KTag	Gloss
ميرسيدس	ETMK•••••SF••••••araDE	Trademark, Mercedes
Amoxillin	EPRO••••••••••••••eng••	Product, medicinal
Aretha Franklin	EFAM•••••SF••••••engUS	Famous person
大手町駅	EFAC•••••S•••••••jpnJP	Facility, Ōtemachi-eki subway station
Ibrahim Traoré	EFAM•••••SM••••••fraBF	Famous person
תורת היחסות הפרטית	ETHR•••••SM••••••hebIL	Theory, Special Relativity
김 정 은	EFAM•••••SM••••••korKP	Famous person, Kim Jong Un
ܓܡܠܐ	ENML•••••SM••••••syr••	Animal, a camel
ⵜⵉⵣⵉ ⵓⵣⵓ	EPPL•••••S•••••••berDZ	City, Tizi Ouzou, Algeria
सगरमाथा	EMRK•••••S•••••••nepNP	Landmark, Mount Everest

If not used for country, cells [21-22] can also be assigned to politeness level; the following examples demonstrate this possibility.

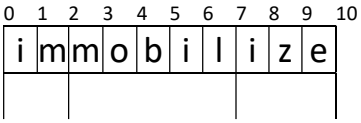
Example	KTag	Gloss
Jacko Wacko	EFAM•••••SM••••••engG•	Famous person, derogatory for Michael Jackson
Japs	EGRP•••••P•••••••engG•	Derogatory for Japanese people

6.2 Morphological analysis examples

Morphological analysis in NLP is the process of breaking down words into their morphemes, and analyzing their grammatical features. It's a crucial step in understanding the structure and meaning of words within a language, enabling various NLP applications. Morphological segmentation can take advantage of KTagset the same way syntactic analysis does.

6.2.1 Morphological segmentation

Generated analysis reports implement only Custom forms which are recommended in morphological KTags, describing individual morph or any character in a surface form is by their position using zero-based indexing meaning the first character in a surface form is at index 0, the second at index 1 and so on. Surface form indexing allows access to individual characters within a morph by their position. Morphemes in the source text file are indexed, each morph is assigned an index incremented by 1; square brackets [] and a colon ":" is used to describe a range of characters representing each morph.



Below are morphologically segmented examples for different languages with morphological KTags shown in the last column; contents in the "Segments" column are for convenience only, not part of this KTagset documentation, demonstrating the boundaries between morphemes and how to index each morpheme correctly.

Language	Example	Segments	Morphological KTag
Amharic			
Arabic	يَعْطِيَانِكُمُوها	يَ.عْطِيَ.انِ.كُمُو.ها	[0:2]FPD?[2:8]VIN[8:11]FSIR?3DM[11:16]FSIR?2PM[16:18]FSIR?3FS
German	arbeitslosigkeit	arbeit.slos.igkeit	[0:6]?[6:10]?[10:16]? unemployment
English	immobilize	im.mobil.ize	[0:2]FPD?[2:7]N[7:10]FSD?
Esperanto	malsanulejo	mal.san.ul.ej.o	[0:3]FPDO?[3:6]FP??[6:8]FSDU[8:10]FSDL[10:11]FSD?
Filipino			
Russian			
Spanish	desflorecer	des.flore.cer	[0:3]FP??[3:8]VI??[8:11]FS??
Swahili			
Thai			
Turkish			
Zulu	ubuntu	u.bu.ntu	[0:1]FPDN[1:3]FP[3:6]N[14]???

KTag should follow the rules below:

- No white spaces for morphological tags
- Indexes are restarted for each surface form
- incorporating extra information like noun classes in Swahili and Zulu

6.2.2 Complex script manipulation

- RTL scripts (Arabic, Hebrew)
- Complex script (Chinese, Korean, Mongolian, Thai)
- No word boundary (Chinese, Japanese)

6.2.3 Diacritical text manipulation

Diacritical text Arabic short vowels "الحركات" and Hebrew Niqqud "נִקּוּד" count when segmenting the surface form.

6.2.4 Representing morphotactics

Morphotactics refers to the set of rules that govern the order in which morphemes can be combined to form words. It's essentially the syntax of morphemes, dictating which morphemes can follow each other and in what order.

Morphologically rich languages implement many tactics take place in the boundaries between morphs of a surface form.

Elaboration: gemination, assimilation, mutation

6.2.5 Morphologically incomplete text

Rich text with diacritical marks occasionally suffers from absence of such important information.

- Tonal marks for languages like Vietnamese and Yoruba,
- Diacriticized characters, dotted, macron, chevron, hyphen etc
- Short vowels in languages like Arabic and Hebrew.

Such situations lead to multiple solutions generated by morphological analyzers; representing all generated solutions using KTagset is possible, the examples below show this feature.

Tonal marks (Vietnamese, Yoruba)

Example	??	KTag

Diacritics

Example	??	KTag

Short vowels (Arabic, Hebrew) ambiguity

Example	??	KTag

6.3 Special use cases for selected languages

Special use cases for some selected languages.

6.3.1 Special use case for Arabic

Arabic morphology and syntax are intertwined systems that govern word formation and sentence structure. Morphology focuses on the internal structure of words and how they are formed from roots and patterns, while syntax deals with how words combine to form phrases and sentences. Arabic is known for its rich morphology through inflection, particularly its verb conjugation and noun declension systems, and for its flexible word order, which can be either VSO or SVO.

Ellipsis, Consonant Mutation, and Vowel Mutation

Example	Segments	Morphological KTag
ازدراد	اِ.زْ.تِ.ر.ا.د	
استيفاء	اِسْتِ.وْ.فْ.ا.ي	
اصطبر	اِ.صْ.تَ.بَر	
السماء	ال.سَمَا.و	
قال	قَ.وَ.لَ	

Dialectal Arabic

Example	Segments	Morphological KTag
كنبغيك	ك.نبغي.ك	
مشينالهم	مشي.نا.لهم	
مرتحتلهاش	م.رتح.ت.ل.ها.ش	

6.3.2 Special use case for Chinese

Chinese, a morphologically "impoverished" language, primarily relies on compounding and derivational morphology rather than inflectional morphology to express grammatical relationships. While it has limited inflectional morphology, Chinese compensates with a rich system of compounding, where two or more morphemes combine to form new words. Chinese syntax, though generally following an SVO structure, has its own unique characteristics, including a tendency towards parataxis and a system of classifiers.

Example	Penn TB	KTag

6.3.3 Special use case for Esperanto

Esperanto has an agglutinative morphology and a relatively free syntax. Words are formed by combining morphemes (roots and affixes), and the language has a relatively simple set of grammatical rules. Its syntax allows for flexible word order due to its case system, but generally follows a (SVO) structure. Esperanto primarily uses affixes to modify word meaning and grammatical function. The language is designed with a high degree of regularity in its morphology, making it relatively easy to learn and predict word formation. Words are formed by combining morphemes (roots, affixes, and endings) without significant changes to the morphemes themselves.

Example	??	KTag

6.3.4 Special use case for Inuktitut

This Inuit language features polysynthetic morphology, allowing for the creation of long, complex words that can express what would be entire sentences in other languages. It uses a system of affixes to convey tense, mood, and other grammatical aspects.

Example	??	KTag

6.3.5 Special use case for Sanskrit

Sanskrit has a highly inflected morphology with a complex system of noun declensions and verb conjugations, including various forms for tense, aspect, and mood. While Sanskrit is generally considered an SOV language, its flexible word order is enabled by its highly inflected nature. Classical Sanskrit also features extensive compounding, which can be analyzed from both syntactic and morphological perspectives.

BIS Tag	KTag	Example	Gloss

BIS POS Tag Malayalam sample

6.3.6 Special use case for Spanish

Spanish syntax generally follows a SVO word order, though it can be more flexible. However, Spanish syntax allows for variations in word order, especially when focusing on specific elements or in specific contexts.

Spanish forms words through inflection, derivation, and compounding. It involves analyzing the internal structure of words, including morphemes and how they combine. Spanish verbs are highly inflected, with many different forms to indicate tense, aspect, mood, person, and number. Nouns and adjectives are also inflected for gender and number; articles and pronouns are inflected as well.

FreeLing Tag	KTag	Example	Gloss

6.3.7 Special use case for Swahili

Swahili is a Bantu language; its syntax generally follows a SVO word order, similar to English, but this can be altered due to the verb's inflection for subject and object agreement, allowing for emphasis. Swahili morphology is characterized by its agglutinative nature, where words are formed by combining morphemes together. It exhibits head-first order in noun phrases and is considered a pro-drop language, meaning subjects and objects can be omitted when recoverable from context. Swahili verbs are highly inflected, with prefixes indicating subject and object agreement, tense, and mood. Nouns are grouped into classes based on prefixes, which also determine agreement patterns with other words in the sentence.

SWATWOL	KTag	Example	Gloss
AD-ADJ	J•••?	kabisa, sana	qualifier of an adjective
ADJ	J•••	zuri, haba	adjective
ADJ-INFL	J•••	zingine	inflecting adjective

6.3.8 Special use case for Tamazight

Tamazight, a Berber language spoken in North Africa, exhibits a complex morphology and syntax, with features common to Afro-Asiatic languages. Its morphology includes extensive apophony (vowel alternations), gender marking on nouns, and templatic structures. Syntactically, Tamazight typically features a VSO word order, although other orders can occur.

BIS Tag	KTag	Example	Gloss

6.3.9 Special use case for Zulu

Zulu is a Bantu language spoken in South Africa, has a rich and complex morphology characterized by its agglutinating nature and a noun class system. This means that words are formed by combining multiple morphemes together, often with prefixes and suffixes. The noun class system, a core feature of Zulu, groups nouns into various classes, each with a distinct prefix, impacting agreement within sentences. Nouns belong to different classes, each marked by a specific prefix. These classes influence agreement patterns with other words in a sentence, like verbs and adjectives.

ZulMorph	KTag	Example	Gloss

Ubuntu u[NPrePre][14]bu[BPre][14]ntu[NStem][14]

Ngiyamkhumbula ngi[SC][1ps]ya[LongPres]mu[OC][1]khumbul[VRoot]a[VT]

6.4 Loanwords consideration

Loanwords can be encoded using KTags by assigning the word with the appropriate KTag and adding the language variety, should be different than the original text.

Example	KTag	Gloss
تسونامي	ENAT••••••••••••••jpn••	Japanese; 津波 Tsunami
Karaoke	N••••••••••••••••jpn••	Japanese; カラオケ
ኦፐራሲዮን	N••••••••••••••••ita••	Italian; Operazione
ガラス	N••••••••••••••••eng••	English; glass
барокко	N••••••••••••••••fra••	French; baroque

6.5 Implementing common tags

Both asterisk “*” and hash “#” replace the inaccessible bullet symbol “•” (U+2022) for testing and quick access; each with slightly different usage.

Tag	Class	Description	Examples
#	Wildcard	Not applicable, place holder	
*	Wildcard	Collective	
?	Unparsed	Unknown words, foreign terms, typo	`asdf`, `lol`, `HTTP/2`
•	Holder	Not applicable, bullet, Unicode U+2022	
∅	Holder	Nothing, U+2205	

7. Mapping different major tagsets

The map tables in this section do not show thorough listing of all possible matching of tags, only the major "universal" categories, to show the potential of KTagset representing every language's morphological and syntactical features. Refer to download section for more details.

Mapping two tagsets should be carried out in context-sensitive semantic manner taking both languages into account that have less far morphologies or at least fall in the same language group; mapping across any pair of different languages requires deep study of the grammatical system of both source and target languages. Examples:

- Arabic "Numerative Noun" matches English "Adverb" for words like (once, twice).
- Hebrew intensive voice matches English adverb, Arabic emphatic present participle. ??

KTags purpose is to only label the grammatical classes, they DO NOT define matching relationships whether syntactical, semantic, or ontological across languages. Suggested mapping should be done for every pair of languages, no general rules to apply on multiple languages or any guarantee that KTagset will fit such purpose except for the major syntactic categories.

7.1 Mapping UniMorph

The goal of UniMorph is to annotate morphological data in a universal schema that allows an inflected word from any language to be defined by its lexical meaning, typically carried by the lemma, and by a rendering of its inflectional form in terms of a bundle of morphological features from the schema.(??)

UniMorph	Description	Equivalent KTag
CONJ	Conjunction	C•••
DET	Determiner	A•••
COMPV	Comparative case	N•C•?
INTJ	Interjection	I•••
INDF	Indefinite	
IPFV	Imperfective	VI••?
INTR	Intransitive verb	V•••••I•••••••••
NOM	Nominative case	N•N•
PROPN	Proper Name	N••N
PRS	Present verb	VI••
PST	Past	VP••
SBJV	Subjunctive mood	V•S•
V	Verb	V•••
V.MSDR	Masdar	NM••
V.PTCP	Participle	

7.2 Mapping Universal Dependencies

The Universal Dependencies (UD) project provides a cross-linguistically consistent tagset with 17 universal POS categories, designed for multilingual parsing and analysis.

UD	Description	Equivalent KTag
ADJ	Adjective	J•••
ADP	Adposition	PA••
ADV	Adverb	D•••
AUX	Auxiliary	
CCONJ	Coordinating conjunction	CC••
DET	Determiner	A•••
INTJ	Interjection	I•••
NOUN	Noun	N•••
NUM	Numeral	M•••
PART	Particle	P•••
PRON	Pronoun	N••P
PROPN	Proper noun	N••N
PUNCT	Punctuation	SP••
SCONJ	Subordinating conjunction	CS••
SYM	Symbol	S•••
VERB	Verb	V•••
X	Other	? •••

7.3 Mapping Penn Treebank

The Penn Treebank tagset remains one of the most widely adopted standards in English NLP applications. It contains 45 distinct tags covering major grammatical categories.

Penn TB	Description	Equivalent KTag
NN	Noun, singular	N••••••••S••••••••••••
NNS	Noun, plural	N••••••••P••••••••••••
NNP	Noun, proper, singular	N••N•••••S••••••••••••
NNPS	Noun, proper, plural	N••N•••••P••••••••••••
VB	Verb, base form	V•••
VBD	Verb, past tense	VP••
VBG	Verb, gerund	
VBN	Verb, past participle	
VBP	Verb, present, non-third person	VI••
VBZ	Verb, present, third person	VI••••••3••••••••••••
JJ	Adjective, base	J•••
JJR	Adjective, comparative	J••C
JJS	Adjective, superlative	J••S
RB	Adverb, base	D•••
RBR	Adverb, comparative	D••C
RBS	Adverb, superlative	D••S

7.4 Mapping Brown Corpus

The Brown Corpus Tagset an earlier but influential tagset containing 87 tags, providing fine-grained distinctions particularly useful for detailed linguistic analysis.

Brown	Description	Equivalent KTag
NN	Noun, singular	N••••••••S••••••••••••
NNS	Noun, plural	N••••••••P••••••••••••
NNP	Noun, proper, singular	N••N•••••S••••••••••••
NNPS	Noun, proper, plural	N••N•••••P••••••••••••
VB	Verb, base form	V•••
VBD	Verb, past tense	VP••
VBG	Verb, gerund	
VBN	Verb, past participle	VP•••10T3DM2PF3SF•••••??
VBP	Verb, present, non-third person	VI••
VBZ	Verb, present, third person	VI••••••3••••••••••••
JJ	Adjective, base	J•••
JJR	Adjective, comparative	J••C
JJS	Adjective, superlative	J••S
RB	Adverb, base	D•••
RBR	Adverb, comparative	D••C
RBS	Adverb, superlative	D••S

7.5 Mapping Bureau of Indian Standard

The Bureau of Indian Standards (BIS) Part of Speech tagset is a standardized set of tags used for annotating parts of speech in Indian languages.

BIS	Description	Equivalent KTag
C	Classifier particle	P•••?
CCD	Coordinator conjunction	CC••
CCS	Subordinator conjunction	CS••
INJ	Interjection	I•••
JJ	Adjective	J•••
NEG	Negation particle	P•••?
NN	Common noun	N•••?
NNP	Proper noun	N••N
NST	Locative inflection	
PRC	Reciprocal pronoun	R••C
PRF	Reflexive pronoun	R••X
PRL	Relative pronoun	R••W
PRP	Personal pronoun	R••U
PRQ	Pronoun question word	R•••?
PUNC	Punctuation	SP••
RB	Adverb	D•••
RDF	Foreign words	
RPD	Default Particle	P•••?
SYM	Symbol	S•••
UNK	Unknown words	?•••
UT	Quotative conjunction	C•••?
VF	Finite verb	V•••?
VINF	Infinitive	N•••?
VNF	Non Finite Verb	V•••?

8. Computational considerations

8.1 KTag composition

User is responsible of composing the correct KTag using custom software agent like *Kalmasoft Verb Conjugation*, *Kalmasoft Noun Declension*. Usage of full length 30 Canonical forms is recommended whenever possible, KTag shortening only when necessary or for demonstration purposes, zero padding for digits and flushing alphabet to the left are best practices for conformity consideration.

8.2 KTag parsing

User is responsible of parsing KTags used in code or corpus using custom parsing agent, this document offers a very basic parser code snippet.

KTags shortening should take into account parser compatibility; in extreme case a parser should practically accept any Custom KTags of any length provided they follow the rules.

8.3 KTag expansion

Expansion renders KTags, if shortened or made Custom, to the Canonical or the closest Short form assuming the place holder character “Bullet, Unicode U+2022” to make for any missing tag element to keep overall uniformity; expansion also “translates” each tag to human readable labels used as glossary for different situations.

Expansion	Example	KTag	Gloss
Short	**ትናገራለህ**	VINA•01T2SM	*You talk*
Expanded	**ትናገራለህ**	VINA•01T2SM••••••	*You talk*; verb-imperfect-indicative-active-form 1-transitive-scond person singular masculine

8.4 KTag translation

The KTagset mapping file provides only the “legal” definitions of each label against its matching tag in the source tagset, a one-to-one relationship of no practical use unless used by a carefully designed algorithm that take every single source tag encountered in the corpus and compose the formal matching KTag that fully describes the same surface form, a process called “translation”.

Context free syntactic analysis yields multiple solutions, morphological analysis mostly generates multiple solutions as well regarding all rich morphology languages.

8.4.1 UniMorph tags translation

Translation of UniMorph feature tags is not straightforward for specific part of speech tags or "features", some may have no matches in KTagset syntax and morphology groups by design e.g. language specific feature "LGSPEC" and verb finiteness "FIN, NFIN", some can be fit only in the semantic and ontology groups of KTagset. UniMorph translation algorithm should basically convert every UniMorph feature to the formal matching KTag considering language specific features. The following examples may clarify the translation goal.

Language	Example	UniMorph	KTag	Gloss
Amharic	ቃሎቻቸው	N;PL;PSS3P;LGSPEC1	N••N•••••P•3P•••• ?	*their words*;
Arabic	كَتَبْتُ	V;PFV;SG;1;MASC	VPNA•01T1SM••••••	*I wrote*;
Armenian	ագահեիք	V;PRF;SBJV;PL;2;FUT		*you would be greedy*;
English	advises	V;3;SG;PRS	VINA•01T3S•••••••	
French	m'accroupissais	V;IND;PST;1;SG;IPFV		*I squatted*;
Fula				
Georgian	შეგშინებოდათ	V;ARGNO2P;IND;PST;PRF	VPNA•••T2P••••••• ?	*would you be scared?*
German	bücher	N;NOM;PL	N••N•••••P•••••••	*books*;
Hausa				
Hebrew				
Hindi				
Japanese				
Korean				
Russian				
Spanish				
Swahili	ninapenda	V;FIN;IND;PRS;1,SG	VINA•01T1SM••••••	*I love*;
Swedish	förhandlades	V;IND;PL;PASS;PST		*was negotiated*;
Tamazight				
Turkish	göndermemişlerdi	V;INFR;PST;PFV;3;PL;NEG;DECL	VPNA••T3P•••••••?	*they didn't send*;
Zulu	bekufika	V;BANTU15;RCT;PST;PROG	VP•A•15I3S•••••••?	*it was coming*;

- Delimiter is not guaranteed to be a semicolon
- Order of feature tags may be arbitrary

8.4.2 Universal Dependency tags translation

Translation of Universal Dependency

Language	Example	UD	KTag	Gloss
Amharic	ቃሎቻቸው		N••N•••••P•3P••••?	*their words*;
Arabic	كَتَبْتُ		VPNA•01T1SM••••••	*I wrote*;
Armenian	ագահեիք			*you would be greedy*;
English	advises		VINA•01T3S•••••••	
French	m'accroupissais			*I squatted*;
Fula				
Georgian	შეგშინებოდათ			*would you be scared?*
German	bücher		N••N•••••P•••••••	*books*;
Hausa				
Hebrew				
Hindi				
Japanese				
Korean				
Russian				
Spanish				
Swahili	ninapenda		VINA•01T1SM••••••	*I love*;
Swedish	förhandlades			*was negotiated*;
Tamazight				
Turkish	göndermemişlerdi		VPNA••T3P•••••••?	*they didn't send*;
Zulu	bekufika		VP•A•15I3S•••••••?	*it was coming*;

8.5 KTagset extension

Lower case Latin letters and Greek alphabet can be used to extend KTagset beyond this document under the sole responsibility of the user.

8.6 Processing considerations

Computationally efficient code needs to carefully consume any resources like main memory and processing power.

8.6.1 KTags shortening

Short form KTags are recommended in applications like corpus automatic annotation, “language variety” cells [18-22] may be removed for monolingual text corpus analysis, Custom KTags may also be used.

In most cases, morphological segmentation can be done using only one index instead of two separated by colon; both KTags below should be acceptable by parsers. The compact version will require careful sequential tracking of indexes to determine the boundaries of each segment.

Language	Example	Version	Morphological KTag
Amharic			
Arabic	يَعْطِيَانِكُمُوها	**Standard**	[0:2]FPD?[2:8]VIN[8:11]FSIR?3DM[11:16]FSIR?2PM[16:18]FSIR?3FS
Arabic	يَعْطِيَانِكُمُوها	**Compact**	[2]FPD?[8]VIN[11]FSIR?3DM[16]FSIR?2PM[18]FSIR?3FS

8.6.2 Large data files manipulation

Tagged corpora usually implement intertwined “morpheme_tag” or “morpheme/tag” style tags that keep the source texts with their assigned tags in one document that is hard to manipulate once generated. KTagset, on the other hand, is designed to be generated separately, ideally as a database or structured file (XML, JSON); a CSV plain text file is also possible provided some tracking mechanism is available to match surface forms to KTags.

File compressing and hashing are suggested techniques to control the data size for efficient information transmission

Tagged corpora for PoS, morphological tags or NER are machine generated files that meant for machine manipulation not human except for demonstration purposes.

- For syntactic analysis and dependency grammar applications, the postag file alone can be used UD??

- For morphological analysis applications, optional punctuation is recommended to parse tokens or surface forms.
- Single source corpus with multiple KTagset solution files, eliminating unnecessary duplication.
- If represented as separate, KTagged file should use same delimiter as the source text file.

8.7 KTagset delivery files

Standard formats for KTagset distribution include:

- **XML**: Structured markup for complex annotations
- **JSON**: Web-friendly format for API integration
- **CoNLL-U**: Universal Dependencies standard format
- **Graph**: ??

9. Future developments

- **Semantic context**: Understanding meaning beyond syntax
- **Ontology context:**
- **Discourse Structure**: Considering document-level patterns
- **Tonal languages**: lacking tone marks like Yoruba
- **Multimodal Information**: Integrating text with other data types

10. Downloads

Resource for download of KTagset related materials.

	File name	Format	Link description
1	KTagset Documentation	PDF	This document
2	Kalmasoft KTagset JSON	JSON	KTagset definition file
3	Kalmasoft KTagset XML	XML	KTagset definition file
4	Kalmasoft KTagset CSV	CSV	
5	UniMorph mapping	JSON	UniMorph to KTagset mapping
6	Universal Dependency mapping	JSON	Universal Depependency to KTagset mapping
7	UniMorph translation	Python	UniMorph to KTags translation code snippet
8	Universal Dependency translation	Python	Universal Depependency to KTag code snippet

11. References

1. UniMorph Tagset: Universal Morphology tagset.
2. Universal Dependencies Tagset: Universal Dependency tagset.
3. HIS Registry of Language Varieties (ROLV): Harvest Information System, Language Varieties.
4. IETF BCP 47: Internet Engineering Task Force Language tag.
5. ISO 639-3: ISO 639-3 Alpha-3 language code.
6. ISO 3166-1: ISO 3166-1 alpha-2 country code.
7. Brown Corpus Tagset: Brown Corpus tagset for the English language.
8. Penn Treebank Tagset: Penn Treebank tagset.
9. BIS Tagset: The Bureau of Indian Standards.
10. AraMorph: AraMorph grammatical categories for the Buckwalter Arabic morphological analyzer.
11. FreeLing Tagset: FreeLing tagset for Spanish.
12. ZulMorph: ZulMorph Tagset for Zulu language morphological analyzer.
13. SWATWOL: Tagset for the Swahili language.
14. Kalmasoft PoS Tagger: Kalmasoft Arabic POS Tagger software.
15. Kalmasoft Arabic Verbs Conjugation: Kalmasoft Arabic verbs conjugation software.
16. Kalmasoft Arabic Nouns Declension: Kalmasoft Arabic nouns declension software.
17. Kalmasoft Arabic Named Entity Recognition: Kalmasoft Arabic NER Extractor software.