International Journal of Information Science and Management
Vol. 16, No. 1, 2018, 165-190

Developing a Comprehensive Standard Persian Positional Tagset

Mohammad Amin Mahdavi
Assistant Professor, Department of Computer Engineering
Imam Khomeini International University, Qazvin, Iran
mahdavi@eng.ikiu.ac.ir

Abstract

One of the primary tools used in text processing tasks such as information retrieval, text extraction, and text mining, is a corpus that is enhnaced by linguistic tags. In a corpus development effort, the role of a POS-tagger is to assign a linguistic tag to every textual token. POS annotation relies heavily on a tagset based on a linguistic theory. Text processing in Persian, too, follows this common practice. Several tagsets have been introduced, so far, to annotate Persian corpora. However, each tagset has followed a specific standard and linguistic theory. The resulting tagsets contain a limited number of tags, which renders them inadequate for a larger scope of research. This study is inspired by EAGLES, MULTEXT-East, positional tagset standards to produce a comprehensive standard positional tagset for Persian. The proposed tagset is also informed by the existing Persian tagsets. The proposed Persian Positional Tagset (PPT) is designed to be used for morphological, lexical, and syntactic annotations of Persian corpora.

Keywords: Persian Positional Tagset, Persian POS tagset, Standard Persian Tagset, Persian Morphosyntactic tagset.

Introduction

A tagged corpus is often used for text processing tasks such as information retrieval, information extraction, and text mining. In a corpus-driven approach, information is retrieved from a corpus that has been enriched with linguistic tags. The annotations in a corpus may include morphological, lexical, semantic and syntactic phenomena. Although it requires more computational resources, extensive annotation in a corpus may yield a richer set of linguistic information. The power to embed a larger amount of information depends on how comprehensive is the tagset with which the corpus is annotated. For the Persian language, too, several tagsets have so far been introduced. However, each of them has served a limited purpose. In this study, we intend to introduce a new comprehensive tagset for Persian that may have the expansion and contraction features to meet various tasks. As part of the initial design requirement, the proposed tagset would have to facilitate morphological, syntactic, and lexical annotations.

The motivations behind this study are two folds. The initial motivation is to introduce a positional tagging scheme for the Persian language; a scheme that has not been presented so far. The additional motivation is to produce a comprehensive set of part-of-speech categories and their respective features for Persian along with the proposed positional tagging scheme.

This study, therefore, intends to propose a comprehensive positional tagset that can be used for morphological, lexical, and syntactic annotation of Persian corpora.

Corpora Development

The development of corpora for different languages is now a well-established tradition among linguists and natural language processing experts. A corpus represents a wide range of language samples. By producing a corpus, one tends to enrich a vast collection of electronic texts using part-of-speech information. To annotate a corpus with part-of-speech information, one would require a tagset, a tagging scheme, and a program that implements the task of tagging (Atwell, 2008).

Some of the pioneering works on corpus development for English language dates back to the Brown corpus (Greene & Rubin, 1971), the Lancaster/Oslo-Bergen corpus (LOB) (Johansson, 1986), Spoken English Corpus (SEC) (Taylor & Knowles, 1988), the Polytechnic of Wales corpus (PoW) (Souter, 1989), the University of Pennsylvania corpus (UPenn) (Santorini, 1990), the London-Lund Corpus (LLC) (Eeg-Olofsson, 1991), the International Corpus of English (ICE) (Greenbaum, 1992, 1993), the British National Corpus (BNC) (Burnard, 2000), and the Spoken Corpus Recordings in British English (SCRIBE) (Huckvale, 2004), among others.

The tradition of developing corpora has continued and also expanded into other languages. Other European languages, East European languages (Dimitrova et al., 1998), Arabic (Khoja, Garside, & Knowles, 2001; Sawalha & Atwell, 2013), and Indo-European languages such as Urdu (Hardie, 2003) have already witnessed extensive tagged corpora. Persian is also one of the Indo-European languages that have followed the corpus development tradition. In the following section, we will explore the corpora projects for Persian.

Persian corpora projects

It is prudent to start with Persian corpora projects to be able to examine part-of-speech tagsets for Persian. As mentioned in the earlier section, every corpus project would require a tagset and a tagging scheme. In this section, an effort will be made to enumerate various Persian corpora projects. Although a few speech corpora have been developed for Persian, this section will only focus on text corpora projects for Persian.

Assi corpus, FLDB

Farsi Linguistics Database (FLDB) is one the earliest corpora initiatives for Persian. The initial version of FLDB contained 3 million words in ASCII format. It was released in Tehran at the Institute for Humanities and Cultural Studies by Mustafa Assi (Assi, 1997). The recent version of the database is in Windows-1256 encoding and it has been renamed to Persian Linguistics Database (PLDB) (Assi, 2005). The newer version includes more than 56 million words. This database is comprised of contemporary literary books, articles, magazines, newspapers, laws and regulations, transcriptions of news, reports, and telephone speeches for lexicography purpose. One advantage of this database is that to each word four linguistic knowledge is attached at once. They are phonetic, syntactic, semantic, and lemma tags. The syntactic tag set contains 44 tags (Assi & Hajiabdolhosseini, 2000).

Hamshahri corpus

Hamshahri is a daily newspaper that keeps an online archive of its issues since 1996. *Hamshahri* has been used as a source for producing Persian corpora by various initiatives. As a dataset for his master's thesis, Ghayoomi has used the *Hamshahri* online archives covering a 6-month period (M Ghayoomi, 2004). This version of the archives contained 6.5 million words.

Darrudi and Hejazi (2004) have employed a crawler to collect available online news from four years of *Hamshahri* archives. After a series of post-processing refinement, they have built a corpus containing 190,206 articles covering politics, city news, economics, reports, editorials, literature, sciences, Society, foreign news, sports, among other categories (Darrudi & Hejazi, 2004). This corpus is comprised of 37 million words.

Shiraz corpus

Shiraz corpus is a parallel corpus that has been tagged for 3,000 Persian sentences and their corresponding translations. The sentences in this corpus have been collected from the online archives of the *Hamshahri* newspaper. All the sentences are manually translated at Computing Research Lab (CRL) of New Mexico State University (Amtrup, Megerdoomian, & Zajac, 2000).

Bijankhan corpus (*Peykareh*)

Perhaps, the most studied Persian corpus to date is that of Bijankhan's, called *Peykareh* (Bijankhan, Sheikhzadeh, Bahrani, & Ghayoomi, 2010). It was developed in 2004. However, it was later revised and expanded to include approximately 38 million words collected from newspapers, books, magazines, articles, technical books, transcription of dialogs, monologues, and speeches for language modeling purpose. The tagging scheme for *Peykareh* is based on the EAGLES standard.

FARSDAT corpus

FARSDATis a corpus developed in 1994 for Persian Speech (Bijankhan & Sheikhzadegan, 1994). The corpus is, in fact, a database of sound files for speech processing and phonetic modeling. The database contains 405 sentence utterances produced by 304 native speakers of Persian. The speakers were selected from ten dialects common in Iran.

Persian language

Persian is an Indo-European language that, in its current form, is written using an adaptation of Arabic script. Persian has been influenced by other languages and, as a result, it has enriched its lexical domain by terms borrowed from other languages.

Persian has a concatenative morphology, in which new morphemes are produced by affixing two or more morphemes. Regarding syntax, Persian has an SOV structure. This project endeavors to capture morphological, lexical, syntactic, and partial semantic information in the proposed tagset. More detailed information on morphology and syntax will be given in later sections when the proposed tagset is explained.

Traditional Persian part of speech categories

The traditional Persian texts on grammar, morphology, and part-of-speech contain an

extensive collection of data on linguistic categories and features. However, one of the leading challenges of the classical works on Persian grammar and part-of-speech is that they lack classification and categorization accuracy. In much of the traditional sources studied for this project, the distinction between lexical categories, syntactic categories, semantic subcategories, and features are often obscured. The focus of this project has been to demarcate main syntactic categories from semantic subcategories and relegate the features away from categories.

Tagsets used in existing Persian corpora

As mentioned in previous sections, several Persian corpora projects have undertaken the task of tagging texts using morphosyntactic tagsets.

In his corpus, Assi introduces a tagset comprising of 43 tags for lexical categories (Assi & Hajiabdolhosseini, 2000). The tags vary from one letter to five letters in length. His tagset follows the criteria proposed by Leech (Assi & Hajiabdolhosseini, 2000). Bijankhan introduced an early version of his tagsets with 550 tags organized in a tree structure (Ghayoomi, Momtazi, & Bijankhan, 2010). Persian Linguistic Database was later released with 44 morphosyntactic tags (Assi, 2005). An interesting Persian-English translation corpus is the Persian translation of the "*1984*" novel by George Orwell, annotated using the MULTEXT-East framework (QasemiZadeh & Rahimi, 2006). Persian Dependency Treebank uses a tagset that, in addition to morphosyntactic annotation, introduces 43 categories for dependency relations (Rasooli, Moloodi, Kouhestani, & Minaei-Bidgoli, 2011).

Standard tagsets for Persian corpora

Several efforts have been made to standardize the development of tagsets for morphosyntactic annotations in Persian corpus projects. Although early tagging initiatives did not adhere to standards, recent projects tend to adopt current standards.

One such standard recommendation was put forward by European Advisory Groups on Language Engineering Standards project (EAGLES). The EAGLES recommendation introduces morphological classes for the European languages (Leech & Wilson, 1999). Bijankhan's *Peykareh* corpus adheres to a tagset that is advised by the EAGLES standard (Bijankhan et al., 2010).

Another initiative that intends to introduce a consistent morphological tagsets for Indo-European languages is MULTEXT-East, which is informed by MULTEXT (Derzhanski & Kotsyba, 2013; Erjavec, 2012). Persian resources are also added to versions 4 and higher of MULTEXT-East. An initial recommendation of MULTEXT-East extension for Persian was introduced by QasemiZadeh and Rahimi (2006).

CLAWS tagger, initially intended for the LOB corpus, was later adopted for other languages such as Urdu (Hardie, 2003) and Arabic (Khoja et al., 2001). Although, unlike MULTEXT and EAGLES, CLAWS is not a standard framework it has been used as a model to develop tagsets for corpus annotations.

Similarly, Penn Treebank tagset is also used as a model to produce a Persian Treebank (Dadegan Research Group (SCICT), 2012; Rasooli et al., 2011; Santorini, 1990). It is noteworthy that all of these standards have been a source of inspiration in developing the tagset proposed in this study.

Describing a positional tagset

Unlike EAGLES and MULTEXT, a positional tagset is not a standardized framework for developing categories and attributes *per se*. It is instead a presentational form, which is a systemic and an easy way to encode and decode morphosyntactic tags. In a positional tagset, each tag is a sequence of characters encoding individual morphological features. In a positional tagset, all tags have the same length. In this system, the position of each character in a tag represents a feature, and the character itself represents the feature value for that position.

In addition to being advised by EAGLES and MULTEXT-East standards, the proposed tagset in this study is also inspired by the positional tagset developed for Russian (Hana & Feldman, 2010) and the Czech positional system (Hajic, 2004). A similar positional tagging system is also used for Arabic (Sawalha & Atwell, 2013).

Introducing Persian positional tagset (PPT)

The proposed tagset, in this study, is intended to cover a wide range of annotations from morphological analysis to Treebank, syntactic, and lexical analysis. Therefore an in-depth study of existing tagsets and standard was considered to ensure that a comprehensive set of tags is achieved. The proposed tagset follows a positional presentational form in which there are 30 positions. In this section, the methodology by which components of PPT have been achieved is also outlined.

Main categories

In terms of the main categories, EAGLES recommendation (Leech & Wilson, 1996) suggests three levels of "obligatory," "recommended," and "optional" constraints. At the "obligatory" level, EAGLES introduces 13 main categories. These categories are Noun (N), Verb (V), Adjective (AJ), Pronoun/Determiner (PD), Article (AT), Adverb (AV), Adposition (AP), Conjunction (C), Numeral (NU), Interjection (I), Unique (U), Residual (R), and Punctuation (PU).

Attribute-value approach to the morphosyntactic framework in EAGLES is also echoed in MULTEXT specification. MULTEXT offers a two-level abstraction (Derzhanski & Kotsyba, 2013; QasemiZadeh & Rahimi, 2006). The first level is the main morphosyntactic categories. MULTEXT specifications are language-specific, and each language has a set of specific categories. However, most Indo-European languages tend to share a lot of common categories. In versions 4 and after, MULTEXT-East specification introduces 12 main categories for Persian. They are Persian Verb (V), Persian Noun (N), Persian Pronoun (P), Persian Determiner (D), Persian Adjective (A), Persian Numeral (M), Persian Adverb (R), Persian Adposition (S), Persian Conjunction (C), Persian Interjection (I), Persian Residual (X), and Persian Abbreviation (Y).

In the positional system, every position denotes an attribute and each letter occupying the position constitutes the value for that attribute (Hana & Feldman, 2010). Therefore, the first position in the positional tag system is reserved for the main categories. The proposed tagset in this study continues to uphold the attribute-value tradition. Similar to EAGLES and MULTEXT-East, Persian Positional Tagset (here we adopt the PPT label for the proposed tagset) reserves the first position for specifying the main categories.

Subcategories

EAGLES specification does not offer more refined subcategories in its obligatory and recommended attributes. Any refinement of the main categories would have to be relegated to optional attributes specific to the language. MULTEXT-East, however, makes the subcategory type more explicit. The positional system is even more vivid in allocating the second position of every tag to the subcategories.

In PPT, subcategories are highlighted as the subtypes of the main categories. Therefore, for every main category, subcategories are outlined in the second position. For obvious reasons, adding an extra level of POS category tends to increase the complexity of taggers. It is also likely that the introduction of subcategories would reduce the efficiency and accuracy of POS taggers. On the other hand, an additional level of granularity allows for semantic and syntactic distinction, which may assist in more refined analysis.

Features

All of the three standards share a common view on an attribute-value pair for features relating to the morphosyntactic categories. Although the set of attributes and their corresponding values are not identical, there is a significant overlap in the attributes and their values reported by MULTEXT-East, EAGLES and positional tag system.

In PPT, additional attributes have been introduced to capture some of the metadata. For instance, knowing whether a constituent, a morpheme, or a lexical item is a stem, derived, inflected, or compound may not be directly relevant to a morphosyntactic category. However, it allows for more refined analysis of the text.

Scope

The proposed tagset is designed for morphological, lexical, and syntactic analyses. For morphological analysis, the proposed tagset would have to be comprehensive enough to cover bound as well as free morphemes. In the case of Persian, the ability to annotate at a morpheme level also assists syntactic analysis. Persian orthography displays some level of fluidity in word demarcation. Sometimes, bound affixes tend to be written detached from their adjacent components. For instance, the plural suffix for nouns (/ha/) can be found separated from the noun itself as in /خانه ها/ (/xaneh/ /ha/). Therefore, it is important to account for such orphan constituent in syntactic analysis.

Part-of-speech taggers are tools that break the text into its lexical tokens. Therefore, for the benefit of tokenization, a tagset would have to be sufficiently detailed to annotate lexical items. Challenges, however, arise when the choices for lexical elements are not clear. Every tagging scheme would have to make assumptions when an ambiguity arises. Persian language, too, has its share of ambiguities. There are multiple lexical items in Persian that give rise to a single token, and *vice versa*. In designing PPT, an effort has been made to capture these lexical phenomena.

At a syntactic level, the proposed PPT tagset has also been designed to annotate phrases and clauses. In PPT, the clause has been treated as the main category. Although it is not part of the traditional POS categories, it is a syntactic category and also a constituent. In terms of the clause being a syntactic category, PPT follows the Penn Treebank categories.

The phrase, however, has not been promoted to the main categories. There is a simple underlying logic that explains this choice. A phrase is a group of tokens that act as a single

category. For instance, a noun phrase replaces a noun. Thus every attribute that is associated with a noun will also be associated with a noun phrase. It is also true at a subcategory level. In other words, a noun phrase can also replace a temporal noun or spatial noun. Therefore, a single attribute value indicating whether the item being tagged is a phrase or a single word would suffice to capture any phrase.

Tag structure

Persian Positional Tags consist of three segments. The first segment is made of one or two uppercase letters representing the main categories. The next segment represented by one or three letters indicate the subcategories for each category. Therefore the first two segments in a tag are reserved for the categories and their respective subcategories. The final segment of the tag is made of positions indicating values for the attributes. A complete list of attributes will be introduced in the following sections. To make the initial segments more readable, we are suggesting a two-letter tag for the main categories and a three-letter tag for the subcategories.

For instance, [NNNOB--------s---sdnic?----?---fp] is a tag that can be assigned to a noun of the object with the following features. The first two uppercase letters indicate the main category (in this case NN=Noun). The following three uppercase letters indicate the subcategory (in this case NOB=Noun of Object). What follows in lowercase letters indicate the values for various features corresponding to the main category, in this case, Noun. Position 9 indicates the feature value for *number* (in this case s = singular). Positions 13 through 18 indicate *specificity* (s = specific), *definiteness* (d = definite), *diminutive* (n = normal), *animate* (i = inanimate), *count* (c = countable), *case* (? = unknown), *adjunct* (? = unknown), *Dependency* (f = free morpheme), and *structure* (p = Primitive Single Root Unit). In the proposed tagset a hyphen (‘-‘) indicates that the feature in that position is not relevant to the current category; hence, not applicable.

Part-of-speech main categories

There are 15 part-of-speech categories proposed in Persian Positional Tagset. Some of these tags are shared with MULTEXT-East and EAGLES tagsets. However, there are some differences between PPT part-of-speech tags an those of MULTEXT-East and EAGLE. These differences are either unique to PPT, or they have been relegated to the subcategory list. Table 1 lists the main categories under PPT tagset.

Table 1

List of main categories for PPT and the corresponding tags in MULTEXT-East and EAGLES. Note that for PPT, two tags are presented. A single letter is given to maintain compatibility with other standard single tags. The two letter tags are to be used as native tags for PPT.

Category Name (Persian)	Category Name (English)	1 LETTER PPT Tag	2 LETTER PPT Tag	MULTEXT Tag	EAGLES Tag
فعل	Verb	V	VB	V	V
اسم	Noun	N	NN	N	N
ضمیر	Pronoun	P	PR	P	PD
مبین	Determiners	T	DT	D	PD

Category Name (Persian)	Category Name (English)	1 LETTER PPT Tag	2 LETTER PPT Tag	MULTEXT Tag	EAGLES Tag
صفت	Adjective	J	AJ	A	AJ
عدد	Numerals	M	NM	M	NU
قید	Adverb	D	AV	R	AV
هم‌نهشت	Adposition	A	AD	S	AP
ربط	Conjunction	C	CN	C	C
وند	Affix	F	AF		
صدا	Interjection	I	IN	I	I
علامت نگارش	Punctuation mark	K	PU		PU
دیگر	Residual	R	RS	X	R
یکتا	Unique	U	UN		U
بند	Clause	S	CL		

Category features

Table 2 describes each of the positions belonging to a category feature. The last 15 columns of the table indicate the relevance of each category feature to the main POS categories.

Table 2

List of all category features based on their position in the tag vector. Feature index is the position in the tag vector.

Feature Index	Feature Name (Persian)	Feature Name (English)	Verb (VB)	Noun (NN)	Pronoun (PR)	Determiners (DT)	Adjective (AJ)	Numerals (NM)	Adverb (AV)	Adposition (AD)	Conjunction (CN)	Affix (AF)	Interjection (IN)	Punctuation mark (PU)	Residual (RS)	Unique (UN)	Clause (CL)
3	شکل	Form	✔														
4	زمان	Tense	✔														
5	نمود	Aspect	✔														
6	مدت	Duration	✔														
7	وجه	mood	✔														
8	شخص	Person	✔		✔												
9	تعداد	Number	✔	✔	✔												
10	ظرفیت	Valence	✔														
11	قطب	Polarity	✔														
12	نهاد	Voice	✔														
13	تعمیم	Specificity		✔													
14	هویت	Definiteness		✔													
15	تصغیر	Diminutive		✔													

Feature Index	Feature Name (Persian)	Feature Name (English)	Verb (VB)	Noun (NN)	Pronoun (PR)	Determiners (DT)	Adjective (AJ)	Numerals (NM)	Adverb (AV)	Adposition (AD)	Conjunction (CN)	Affix (AF)	Interjection (IN)	Punctuation mark (PU)	Residual (RS)	Unique (UN)	Clause (CL)
16	حرکت	Animacy		✔													
17	شمار	Count		✔													
18	حالت	Case		✔	✔												
19	واژبست	Clitic	✔	✔	✔	✔	✔										
20	شخص و شماره‌ی واژبست	Clitic number and person	✔	✔	✔	✔	✔										
21	ادب	Courtesy	✔		✔												
22	موقعیت	Position				✔	✔	✔		✔							
23	نماد	Representation					✔	✔									
24	شدت	Degree					✔		✔								
25	متمم	Adjunct		✔	✔	✔	✔	✔	✔	✔	✔						✔
26	نوع ربط	Conjunction Type									✔						
27	نوع وند	Affix Type										✔					
28	نوع علامت	Punctuation Type												✔			
29	وابستگی	Dependency	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔		✔	✔
30	ساختار	Structure	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔		✔	✔

Common Features

Persian positional tagset identifies 28 features in total that make up the feature vector for morphosyntactic tags. While some of these features are peculiar to one or more categories, two features are common to all categories and apply to every tag regardless of its morphosyntactic category.

Table 3

List of common category features and their respective possible feature values.

Vector Index	Feature (English)	Feature (Persian)	Feature Value (English)	Feature Value (Persian)	3 Letter PPT Tag	1 Letter PPT tag
29	Dependency	وابستگی	Unknown	مجهول	UKN	?
29	Dependency	وابستگی	Not Applicable	نامربوط	IRR	-
29	Dependency	وابستگی	Unchanged	بلاتغییر	UCH	=
29	Dependency	وابستگی	Free	آزاد	FRE	f
29	Dependency	وابستگی	Bound	مقید	BND	b
30	Structure	ساختار	Unknown	مجهول	UKN	?
30	Structure	ساختار	Not Applicable	نامربوط	IRR	-
30	Structure	ساختار	Unchanged	بلاتغییر	UCH	=

Vector Index	Feature (English)	Feature (Persian)	Feature Value (English)	Feature Value (Persian)	3 Letter PPT Tag	1 Letter PPT tag
30	Structure	ساختار	Primitive Single Root Unit	تکواژ جامد تک ریشه	PRM	p
30	Structure	ساختار	Derived Single Root Unit	تکواژ مشتق تک ریشه	DRV	d
30	Structure	ساختار	Inflicted Single Root Unit	تکواژ صرفی تک ریشه	NFL	f
30	Structure	ساختار	Compound Multi Root Unit	تکواژ مرکب چند ریشه	CMP	c
30	Structure	ساختار	Hybrid Lexical Unit	تکواژ مختلط	HYB	h
30	Structure	ساختار	Orphan Word	یتیم‌واژه	ORP	o
30	Structure	ساختار	Multi-Word Lexical Unit	پاره‌واژ	MWU	m
30	Structure	ساختار	Phrase	گروه	PHR	p
30	Structure	ساختار	Clause	بند	CLS	s

Feature values

Table 4
List of possible values for each of the positions in the category feature vector.

Feature Index	Feature Name (Persian)	Feature Name (English)	Feature Value (Persian)	Feature Value (English)	Three Letter Tag	Single Letter Tag
3	شکل	Form	ماده	Participle	PPL	p
3	شکل	Form	ستاک	Stem	STM	s
3	شکل	Form	مصدر	Infinitive	INF	i
3	شکل	Form	صرفی	Conjugated	CNG	c
4	زمان	Tense	آینده	Future	FUT	f
4	زمان	Tense	حال	Present	PRS	r
4	زمان	Tense	گذشته	Past	PAS	p
5	نمود	Aspect	پیشین نقلی (نقلی بعید)	Precedent Narrative	PNR	r
5	نمود	Aspect	پیشین (بعید)	Precedent	PRC	c
5	نمود	Aspect	نقلی (بازگویی)	Narrative	NAR	n
5	نمود	Aspect	ساده	Simple	SMP	s
6	مدت	Duration	استمراری	Progressive	PRG	g
6	مدت	Duration	ناتمام	imperfect	IPF	i
6	مدت	Duration	تمام	perfect	PRF	p
7	وجه	mood	آرزوئی	Optative	OPT	o
7	وجه	mood	گُمانی	Presumptive	PMT	s
7	وجه	mood	شرطی	Potential (conditional)	POT	p
7	وجه	mood	درخواستی (امری)	Imperative	IMP	m
7	وجه	mood	تردیدی (التزامی)	Subjunctive	SUB	s
7	وجه	mood	گزارشی (اخباری)	Indicative	IND	i
8	شخص	Person	سوم شخص	Third	3RD	3

Feature Index	Feature Name (Persian)	Feature Name (English)	Feature Value (Persian)	Feature Value (English)	Three Letter Tag	Single Letter Tag
8	شخص	Person	دوم شخص	Second	2ND	2
8	شخص	Person	اول شخص	First	1ST	1
9	تعداد	Number	گروه	Collective	COL	c
9	تعداد	Number	مثنی	Dual	DUL	d
9	تعداد	Number	جمع مکسر	Irregular Plural	IPL	i
9	تعداد	Number	جمع سالم	Sound Plural	PLR	p
9	تعداد	Number	مفرد	Singular	SNG	s
10	ظرفیت	Valence	دو مفعولی	Ditransitive	DTR	d
10	ظرفیت	Valence	متعدی	Transitive	TRN	t
11	قطب	Polarity	منفی	Negative	NEG	n
11	قطب	Polarity	مثبت	Positive	POS	p
12	نهاد	Voice	مجهول	Passive	PSS	p
12	نهاد	Voice	معلوم	Active	ACT	a
13	تعمیم	Specificity	عام	Non-Specific	NSP	n
13	تعمیم	Specificity	خاص	Specific	SPC	s
14	هویت	Definiteness	نکره	Indefinite	IDF	i
14	هویت	Definiteness	معرفه	Definite	DEF	d
15	تصغیر	Diminutive	هنجار	Normal	NOR	n
15	تصغیر	Diminutive	مصغر	Diminutive	DIM	d
16	حرکت	Animacy	ساکن	Inanimate	INA	i
16	حرکت	Animacy	متحرک	Animate	ANM	a
17	شمار	Count	ناشمردنی	Mass	MAS	m
17	شمار	Count	شمردنی	Count	CNT	c
18	حالت	Case	وجودی	Essive	ESS	j
18	حالت	Case	عدمی	Abessive	ABV	f
18	حالت	Case	همراهی	Comatative	CMT	o
18	حالت	Case	پایانی	Terminative	TER	m
18	حالت	Case	تحولی	Translative	TRS	r
18	حالت	Case	بخشی	Partitive	PRT	p
18	حالت	Case	ندایی	Vocative	VOC	v
18	حالت	Case	مفعولی (غیر مستقیم، مکان)	Allative (going to a place)	ALL	e
18	حالت	Case	مفعولی (غیر مستقیم، مکان)	Elative	ELV	a
18	حالت	Case	مفعولی(غیر مستقیم، مکان)	Adessive (being near a place)	ADS	s
18	حالت	Case	مفعولی (غیر مستقیم، مکان)	Illative (entering a place)	ILL	t

Feature Index	Feature Name (Persian)	Feature Name (English)	Feature Value (Persian)	Feature Value (English)	Three Letter Tag	Single Letter Tag
18	حالت	Case	مفعولی (غیر مستقیم، مکان)	Inessive	INE	i
18	حالت	Case	مفعولی (غیر مستقیم، مکان)	Locative (Pure location)	LOC	l
18	حالت	Case	مفعولی (مستقیم)	Accusative	ACC	c
18	حالت	Case	مفعولی (غیر مستقیم)	Ablative	ABL	a
18	حالت	Case	مفعولی (غیر مستقیم)	Oblique	OBL	b
18	حالت	Case	مفعولی (غیر مستقیم)	Dative	DAT	d
18	حالت	Case	اضافی	Genitive	GEN	g
18	حالت	Case	نهادی	Nominative	NOM	n
19	واژبست	Clitic	هم‌نهشت	Adposition	ADP	d
19	واژبست	Clitic	ضمیر	Pronominal	PRN	p
19	واژبست	Clitic	صفت	Adjectival	AJL	a
19	واژبست	Clitic	اسم	Nominal	NML	n
19	واژبست	Clitic	عطف	Conjunction	CNJ	k
19	واژبست	Clitic	مبین	Determiner	DTM	r
19	واژبست	Clitic	یای اضافه	Genitive 'ye'	EZY	y
19	واژبست	Clitic	کسره اضافه	Genitive 'e'	EZE	e
20	شخص و شماره‌ی واژبست	Clitic number and person	سوم جمع	Third Plural	3PL	6
20	شخص و شماره‌ی واژبست	Clitic number and person	دوم جمع	Second Plural	2PL	5
20	شخص و شماره‌ی واژبست	Clitic number and person	اول جمع	First Plural	1PL	4
20	شخص و شماره‌ی واژبست	Clitic number and person	سوم مفرد	Third Singular	3SG	3
20	شخص و شماره‌ی واژبست	Clitic number and person	دوم مفرد	Second Singular	2SG	2
20	شخص و شماره‌ی واژبست	Clitic number and person	اول مفرد	First Singular	1SG	1
21	ادب	Courtesy	خودمانی	Familiar	FML	f
21	ادب	Courtesy	مؤدب	Polite	PLT	p
22	موقعیت	Position	هردو	Both	BTH	b
22	موقعیت	Position	پسین	Suffixed	PST	s
22	موقعیت	Position	پیشین	Prefixed	PRE	p
23	نماد	Representation	مخلوط	Mix	MIX	m

Feature Index	Feature Name (Persian)	Feature Name (English)	Feature Value (Persian)	Feature Value (English)	Three Letter Tag	Single Letter Tag
23	نماد	Representation	حروف	Letter	LET	l
23	نماد	Representation	رومی	Roman	ROM	r
23	نماد	Representation	رقم	Digit	DIG	d
24	شدت	Degree	تفضیلی	Arabic Superlative	ASP	a
24	شدت	Degree	مبالغه (عربی)	Arabic Elative	ELT	e
24	شدت	Degree	برترین	Superlative	SUP	s
24	شدت	Degree	برتر	Comparative	CMP	c
24	شدت	Degree	مطلق	Absolute	ABS	s
25	متمم	Adjunct	خیر	No	NOT	n
25	متمم	Adjunct	بلی	Yes	YES	y
26	نوع ربط	Conjunction Type	پیرو	Subordination	SBR	s
26	نوع ربط	Conjunction Type	عطفی	Coordination	COD	c
27	نوع وند	Affix Type	صرفی	Inflectional	IFL	i
27	نوع وند	Affix Type	اشتقاقی	Derivational	DER	d
28	نوع علامت	Punctuation Type	ویژه	Special	SPL	s
28	نوع علامت	Punctuation Type	گروه‌ساز	Grouping	GRP	g
28	نوع علامت	Punctuation Type	جداساز	Separator	SEP	p
28	نوع علامت	Punctuation Type	جمله	Sentence	SNT	n
29	وابستگی	Dependency	مقید	Bound	BND	b
29	وابستگی	Dependency	آزاد	Free	FRE	f
30	ساختار	Structure	بند	Clause	CLS	s
30	ساختار	Structure	گروه	Phrase	PHR	p
30	ساختار	Structure	پاره‌واژ	Multiword Lexical Unit	MWU	m
30	ساختار	Structure	یتیم‌واژه	Orphan Word	ORP	o
30	ساختار	Structure	تکواژ مختلط	Hybrid Lexical Unit	HYB	h
30	ساختار	Structure	تکواژ مرکب چند ریشه	Compound Multi Root Unit	COM	c
30	ساختار	Structure	تکواژ صرفی تک ریشه	Inflicted Single Root Unit	NFL	f
30	ساختار	Structure	تکواژ مشتق تک ریشه	Derived Single Root Unit	DRV	d
30	ساختار	Structure	تکواژ جامد	Primitive Unit	PRM	p

Common values

Similar to the features that are common to all categories, several feature-values are

common to all features. The obvious values for all features are "Unknown" and "Not Applicable." Before the task of POS tagging is begun, all of the feature values may be set to "Unknown." As soon as the main category is identified, all features irrelevant to that category may be set to "Not Applicable." There is a third common value that has been assign to each feature that facilitates the disambiguation task during the POS tagging process. The third common feature value is "Unchanged," which indicated the lack of value transformation.

Table 5
List of common feature values that appear for all category features

Common Feature Value (Persian)	Common Feature Value (English)	3 Letter PPT Tag	1 Letter PPT tag
مجهول	Unknown	UKN	?
نامربوط	Not Applicable	IRR	-
بلاتغییر	Unchanged	UCH	=

Subcategories

The following 15 subsections outline the POS subcategories for each main category. These subcategories are syntactic as well as semantic in nature.

Verb subcategories

Table 6
List of subcategories for VERB.

Verb Subcategory (Persian)	Verb Subcategory (English)	3Letter PPT Tag	1 Letter PPT tag
فعل کنشی	Action verb	ACT	a
فعل ربطی	Copula	COP	c
فعل کمکی	Auxiliary	AUX	u
فعل شبه کمکی	Modal	MOD	m
فعل ضعیف	Light verb	LGT	l

Noun subcategories

Table 7
List of subcategories for NOUN.

Noun Subcategory (Persian)	Noun Subcategory (English)	3Letter PPT Tag	1 Letter PPT tag
اسم عنصر(مواد سازنده یا جنس)	Noun of Element	ELM	e
اسم اشیاء یا محصول	Noun of Object or Product	NOB	n
اسم خصوصیت	Noun of Trait	TRT	t
اسم انتزاعی (معنی)	Abstract Noun	ABS	a
اسم شخص	Personal Noun	PSN	p
اسم کمپانی یا شرکت	Business Name	BSN	b
نام تجاری	Brand Name	BRN	r
اسم حیوان	Animal Name	ANL	m

Noun Subcategory (Persian)	Noun Subcategory (English)	3Letter PPT Tag	1 Letter PPT tag
اسم پرنده	Bird Name	BIR	d
اسم گیاه	Plant Name	PLN	t
اسم ابزار	Instrumental Noun	INS	i
اسم واقعه	Noun of Event	EVN	v
اسم نقطه‌ی زمان	Temporal Noun (Point)	TPN	o
اسم بازه‌ی زمان	Temporal Noun (Period)	TPR	s
اسم مدت زمان	Temporal Noun (Duration)	TDR	u
تاریخ	Date	DAT	t
اسم مکان	Spatial Noun	SPT	y
اسم مکان جغرافیایی	Geographic Spatial Noun	GEO	g
اسم فاعل (ماده‌ی حال)	Active Participle	APP	k
اسم فاعل حرفه	Active Participle Profession	APJ	h
اسم مفعولی (ماده گذشته)	Passive Participle	PPP	c
اسم مبالغه	Elative Noun	ELN	f
اسم صفتی	Adjectival Noun	AJN	j
مصدر فارسی	Persian Infinitive verb (Gerund)	GRP	l
مصدر عربی	Arabic Infinitive verb (Gerund)	GRA	q
اسم مصدر	Nominalized Infinitive	NMI	w
اسم فعل	Verbal Noun	VRN	z
مصدر مرخم	Reduced Infinitive (Gerund)	GRN	$
عدد	Numeral	NMR	#

Pronoun subcategories

Table 8
List of subcategories for PRONOUN.

Pronoun Subcategory (Persian)	Pronoun Subcategory (English)	3Letter PPT Tag	1 Letter PPT tag
ضمیرِ شخصی منفصل (فاعلی و مفعولی)	Personal Pronouns (subject and object)	PSP	p
ضمیر فاعلی حال	Subject Present	SBP	s
ضمیر فاعلی گذشته	Subject Past	PST	t
ضمیر فاعلی التزامی	Subjunctive	SBJ	b
ضمیر مفعولی	Object Pronoun	OBP	o
ضمیرِ دارائی	Possessive Pronouns	POS	v
ضمیرِ بازتابی و کانونی	Reflexive and Intensive Pronouns	RFP	r
ضمیرِ تمایزی	Differential Pronouns	DFP	d
ضمیرِ رو به روئی	Reciprocal Pronouns	RCP	c
ضمیرِ اشاره	Demonstrative Pronouns	DMP	v
ضمیرِ پرسشی	Interrogative Pronouns	INP	g

Pronoun Subcategory (Persian)	Pronoun Subcategory (English)	3Letter PPT Tag	1 Letter PPT tag
ضمیرِ پخشی (توزیعی)	Distributive Pronouns	DSP	u
ضمیرِ همگانی	Associative Pronouns	ASP	a
ضمیرِ تهی	Non-existential Pronouns	NXP	i
ضمیرِ گزینشی	Selective Pronouns	SEL	l
ضمیرِ بخشی	Partitive Pronouns	PRP	e
ضمیرِ چگونگی	Qualifying Pronouns	QLP	q
ضمیرِ اندازه	Quantifying Pronouns	QTP	y
ضمیرِ شگفتی	Exclamatory Pronouns	EXP	m
ضمیرِ تکمیلی	Elative Pronouns	ELP	n
ضمیرِ نسبی	Multiplicative Pronouns	MUL	h
ضمیرِ شمارشی	Counting Pronouns	NMP	#

Determiner subcategories

It must be highlighted here that PPT also introduces the "direct object identifier" ('ra') as a determiner rather than an Adposition.

Table 9

List of subcategories for DETERMINER.

Determiner Subcategory (Persian)	Determiner Subcategory (English)	3Letter PPT Tag	1Letter PPT Tag
مبین بازتابی و کانونی	Reflexive and Intensive Determiner	RFD	r
مبین تمایزی	Differential Determiner	DFD	f
مبین اشاره	Demonstrative Determiner	DMD	d
مبین پرسشی	Interrogative Determiner	IND	i
مبین پخشی	Distributive Determiner	DSD	s
مبین همگانی	Associative Determiner	ASD	a
مبین تهی	Non-existential Determiner	IXD	e
مبین بخشی	Partitive Determiner	PRD	p
مبین چگونگی	Qualifying Determiner	QLD	q
مبین اندازه	Quantifying Determiner	QTD	n
مبین شگفتی	Exclamatory Determiner	EXD	l
مبین تکمیلی	Elative Determiner	ELD	t
مبین شمارشی	Counting Determiner	NMD	#
مبین آماری	Ordinal Determiner	ORN	o
مبین نکره	Indefinite Article	IDD	u
مبین مفعولی	Accusative Particle	ACC	c
مبین اضافی	Genitive Particle	EZF	z
مبین اضافی	Genitive Particle	YZF	y

Adjective subcategories

Table 10
List of subcategories for ADJECTIVE.

Adjective Subcategory (Persian)	Adjective Subcategory (English)	3Letter PPT Tag	1 Letter PPT tag
بیانی ساده	Description	SMP	d
بیانی فاعلی	Subject	SUB	s
بیانی مفعولی	Object	OBJ	o
بیانی نسبی	Relation	REL	r
بیانی لیاقت	Deserving	DES	s
بیانی رنگ	Color	CLR	c
بیانی شکل	Shape	SHP	h
بیانی جنس	Material	MAT	m
بیانی وضع	State	STA	t
بیانی کیفی	Quality	QLJ	q
بیانی مقدار	Quantity	QTJ	n
بیانی اندازه	Size	SIZ	z
بیانی وزن	Weight	WGH	w
بیانی شخصیت	personality	PRS	p
بیانی احساس	emotion	EMO	e
بیانی صدا	Sound	SND	u
بیانی طعم	Taste	TST	f
بیانی لمس	Touch	TCH	k
بیانی سرعت	Speed	SPD	i
بیانی دما	Temperature	TMP	j
بیانی سن	Age	AGE	g
بیانی روشنایی	Brightness	BGH	u
بیانی زمان	Time	TIM	v
بیانی مکان	Location	LOC	l
بیانی فاصله	Distance	DST	w
بیانی بویایی	Smell	SML	y
شمارشی آماری	Cardinal Numerals	CRJ	#
شمارشی ترتیبی	Ordinal Numerals	ORJ	&

Numeral subcategories

Table 11
List of subcategories for NUMERAL.

Numeral Subcategory (Persian)	Numeral Subcategory (English)	3Letter PPT Tag	1 Letter PPT tag
شماره‌ی آماری	Cardinal Numerals	CRN	c
شماره‌ی ترکیبی	Compound Numerals	CMN	M
شماره‌ی ترتیبی	Ordinal Numerals	ORN	o
شماره‌ی بخشی	Partitive Numerals	PRN	p
شماره‌ی کسری	Fractional Numerals	FRC	f
شماره‌ی اعشاری	Decimal Numerals	DCM	d
شماره‌ی توزیعی	Distributive Numerals	DSN	s
واحد سنجش	Measure Word	MSR	z
واحد شمارش	Counting Unit	UNT	u
واحد گروه	Grouping Unit	GRP	g

Adverb subcategories

Table 12
List of subcategories for ADVERB.

Adverb Subcategory (Persian)	Adverb Subcategory (English)	3 Letter PPT Tag	1 Letter PPT Tag
قید اختصار	Summarization	SMZ	$
قید علت	Cause	CUS	@
قید تصدیق و تأکید	Affirmative	AFF	a
قید تشبیه	Simile	SIM	b
قید تداوم	Continuity	CON	c
قید ترتیب	Order	ORV	d
قید استثنا	Exception	EXT	e
قید تکرار	Frequency	FRQ	f
قید پرسش	Interrogative	ING	g
قید انحصار	Exclusivity	EXS	h
قید قصد	Intention	INT	i
قید تردید	Subjunctive	SJC	j
قید سوگند	Oath	OTH	k
قید کیفیت (چگونگی)	Quality	QLV	l
قید حالت	Manner	MAN	m
قید نفی	Negation	NEG	n
قید آرزویی	Optative	OPT	o
قید تفسیر	Interpretive	INR	p
قید مقدار (کمیت)	Quantity	QTV	q

Adverb Subcategory (Persian)	Adverb Subcategory (English)	3 Letter PPT Tag	1 Letter PPT Tag
قید تأسف	Regret	REG	r
قید مکان	Spatial	SPC	s
قید زمان	Temporal	TMV	t
قید اصرار	Accentual	ACN	u
قید تبری و ادب	Forgiveness	FGV	v
قید شرط	Conditional	CND	w
قید تعجب	Exclamation	EXC	x
قید استعلا	Transcend	TRN	y
قید اندازه	Measure	MSU	z

Adposition subcategories

Persian seems to have prepositions only. Although some sources (Bijankhan et al., 2010) report just one postposition for Persian (\ra\), PPT recognizes that postposition as a determiner for the direct object. The postposition reported in Bijankhan et al. is a marker for the direct object (‘ra’). Since it acts as a constant marker for the direct object, PPT relegates this marker to the family of determiners which accompany nouns and noun phrases at all times.

Table 13
List of subcategories for ADPOSITION.

Adposition Subcategory (Persian)	Adposition Subcategory (English)	3Letter PPT Tag	1 Letter PPT tag
هم‌نهشت شروع زمان	Adposition of Starting Time	TMS	s
هم‌نهشت پایان زمان	Adposition of Ending Time	TME	e
هم‌نهشت فاصله‌ی زمان	Adposition of Time Duration	TMD	l
هم‌نهشت فاصله‌ی مکان	Adposition of Distance	PLD	d
هم‌نهشت جهت	Adposition of Direction	DIR	r
هم‌نهشت همراهی	Adposition of Companion	ACM	c
هم‌نهشت نهادی	Adposition of Agent	AGN	g
هم‌نهشت ابزاری	Adposition of Instrument	INA	i
هم‌نهشت منظور	Adposition of purpose	PUR	p
هم‌نهشت اندازه	Adposition of Measure	MSA	z
هم‌نهشت حالت	Adposition of State	ADS	t
هم‌نهشت شرط	Adposition of Condition	CDN	k
هم‌نهشت اعترافی	Adposition of Concession	CNS	o

Conjunction subcategories

Table 14
List of subcategories for CONJUNCTION.

Conjunction Subcategory (Persian)	Conjunction Subcategory (English)	3Letter PPT Tag	1 Letter PPT tag
همپایه	Copulative	CPL	c
نقیض	Adversative	ADV	a
گزینه	Alternative	ALT	l
پیشرفت	Order	ORC	o
پیرو مکان	Subordination of Place	SPL	p
پیرو زمان	Subordination of Time	STM	t
پیرو چگونگی	Subordination of Manner	SMN	m
پیرو دلیل	Subordination of Reason	SRN	r
پیرو قیاس	Subordination of Comparison	SCM	i
پیرو منظور	Subordination of Purpose	SPR	s
پیرو احتمال	Subordination of mode	PRB	d
پیرو نتیجه	Subordination of Result	SRS	u
پیرو شرط	Subordination of Condition	SCN	n
پیرو توجیه	Subordination of Substantiation	SSB	b

Affix subcategories

Affixes are either prefix or suffix. The subcategories are further divided into affixes and suffixes that contribute to the formation of a noun, an adjective, an adverb, or a verb. One commonly case of agreed upon infix in Persian is that of "/ا/." Therefore, infix is not divided into further divisions.

Table 15
List of subcategories for AFFIX.

Affix Subcategory (Persian)	Affix Subcategory (English)	3Letter PPT Tag	1 Letter PPT tag
پیشوند اسمی	Noun Prefix	PXN	n
پیشوند فعلی	Verb Prefix	PXB	v
پیشوند صفتی	Adjective Prefix	PXJ	j
پیشوند قیدی	Adverb Prefix	PXV	d
پسوند اسمی	Noun Suffix	SXN	f
پسوند فعلی	Verb Suffix	SXV	b
پسوند صفتی	Adjective Suffix	SXJ	c
پسوند قیدی	Adverb Suffix	SXD	d
میانوند	Infix	INF	i

Interjection subcategories

Table 16
List of categories for INTERJECTION.

Interjection Subcategory (Persian)	Interjection Subcategory (English)	3Letter PPT Tag	1 Letter PPT tag
تحسین و آفرین	Admiration	ADM	a
شگفتی و تعجب	Amazement	AMZ	m
تنبیه	Admonishment	ADN	d
تأسف و دریغ	Contrition	CTR	c
نفرت و بیزاری	Abomination	ABM	b
اظهار درد	Lamentation	LAM	l
ندا	Vocative	VOC	v
استهزا و ریشخند و تحقیر	Diminutive	DIM	i
تفریح و شادی و خوشایند بودن	Delectation	DLC	e
پناه بردن	Appeal	APL	p
ترساندن و پرهیز دادن	Intimidation	INM	t
دعا	Invocation	INV	n

Punctuation subcategories

Table 17
List of subcategories for PUNCTUATION.

Punctuation Subcategory (Persian)	Punctuation Subcategory (English)	3Letter PPT Tag	1 Letter PPT tag
نقطهٔ	Full stop	STP	f
ویرگول	Comma	COM	c
دو نقطه	Colon	CLN	n
نقطه ویرگول	Semicolon	SMI	s
پرانتز چپ	Left Parentheses	PRL	l
پرانتز راست	Right Parentheses	PRR	r
کروشه چپ	Left Square Brackets	SQL	p
کروشه راست	Right Square Brackets	SQR	q
کوچک‌تر	Less Than	LTH	t
بزرگ‌تر	Greater Than	GTH	u
علامت نقل قول یگانه	Single Quotation Mark	QTS	a
علامت نقل قول دوگانه	Double Quotation Mark	DQT	d
تیره	Hyphen	HYP	h
سؤال	Question mark	QUE	m
تعجب	Exclamation mark	EXM	x
حذف	Ellipsis mark	ELL	i

Punctuation Subcategory (Persian)	Punctuation Subcategory (English)	3Letter PPT Tag	1 Letter PPT tag
تساوی	Equal sign	EQL	e
آکولاد چپ	Left Curly Brackets	CUL	v
آکولاد راست	Right Curly Brackets	CUR	w
درصد	Percent Sign	PER	g
مورب	Slash	SLA	j

Residual subcategories

The residual category is applied to all elements that fall within the miscellaneous groups. While this category may grow in tagging projects with a specific task, for the general tagging purposes, Table 26 should cover the common elements.

Table 18
List of subcategories for RESIDUAL.

Residual Subcategory (Persian)	Residual Subcategory (English)	3Letter PPT Tag	1 Letter PPT tag
مخفف	Abbreviation	ABR	b
کدواژ	Acronym	ACR	c
ایمیل	Email	EML	e
تارنما	URL	URL	u
شماره‌ی تلفن	Telephone	TEL	t
آدرس	Address	ADD	a
واژه‌ی بیگانه	Foreign Word	FGN	f
فرمول	Formula	FML	m
ناشناخته	Unknown	UNK	k

Unique subcategories

Unique is a category recommended by MULTEXT to cover subcategories that are unique to particular languages. In PPT, "unique" was chosen to identify single word elements that are made up of more than one POS. Other tagging schemes refer to this type of element as MUT (Multi-Unit Token) (Sharifi-Atashgah & Bijankhan, 2009). However, the terminology adopted for it in PPT is "Hybrid." These elements are easily distinguished from "phrases" since they are not multi-words.

There is one more unique subcategory introduced in PPT that relates to the absence of an element. This subcategory is called a placeholder. One obvious such case is the absence of the subject pronoun from the sentence. Since the inflected form of the verb implicitly indicates the subject pronoun, the explicit presence of subject pronoun is not necessary. This tag may assist the anaphora resolution tasks in discourse analysis.

Table 19
List of subcategories for UNIQUE.

Unique Subcategory (Persian)	Unique Subcategory (English)	3Letter PPT Tag	1 Letter PPT tag
تکواژ مختلط گروه اسمی	Hybrid Morpheme Noun Phrase	HNP	n
تکواژ مختلط گروه صفتی	Hybrid Morpheme Adjectival Phrase	HAP	j
تکواژ مختلط گروه اسنادی صفتی	Hybrid Morpheme Predicative Adjectival Phrase	HPJ	p
تکواژ مختلط گروه اسنادی اسمی	Hybrid Morpheme Predicative Noun Phrase	HPN	h
تکواژ مختلط گروه فعلی	Hybrid Morpheme Verb Phrase	HVP	v
جانما	Placeholder for Absent Constituent	PLC	b

Clause subcategories

Table 20
List of subcategories for CLAUSE

Clause Subcategory (Persian)	*Clause Subcategory (English)*	*3Letter PPT Tag*	*1 Letter PPT tag*
جمله‌ی خبری	Declarative Sentence	DCL	c
جمله‌ی خبری تعجبی	Exclamatory Sentence	EXL	e
جمله‌ی پرسشی قطبی	Polar Interrogative Sentence	PIS	p
جمله‌ی پرسشی عطفی	Disjunctive Interrogative Sentence	DIS	D
جمله‌ی پرسشی ادواتی	Constituent Interrogative Sentence	CIS	i
جمله‌ی امری	Imperative Sentence	IMS	m
جمله‌ی آرزویی	Optative Sentence	OPS	o
جمله‌ی پیرو	Subordinate Relative Clause	SRC	r
جمله‌ی پیرو متمم	Subordinate Adverbial Clause	SAC	S

Concluding remarks

Several corpora have been produced for Persian over the years. Each corpus has used a specific tagset to serve a limited purpose. It has yielded several non-compatible tag systems for the Persian language. Most of the tags developed so far are used for syntactic part-of-speech tagging at the lexical level. However, recent efforts have been made to produce treebanks and fine-grained morphological analysis. What has inspired this study is the need for the development of the standard tagset that may be used for morphological, lexical, and syntactic tagging.

In this study, we have been able to propose a positional tagset for Persian that follows tagset standards of MULTEXT and EAGLES. This tagset consists of 15 main categories that are further subdivided into syntactic and semantic subcategories. The corresponding features are laid out in a vector that is made up of 30 positions. The scope of the Persian Positional Tagset covers morphological, lexical, and syntactic phenomena.

One of the advantages of the proposed tagset is its ability to facilitate partial tagging of

linguistic elements. The existing tagsets are made up of a fixed number of morphosyntactic tags. Most of them are used for high-level part of speech tagging. This shallow tagging tends to overlook more detailed linguistic information captured by various feature values. In other words, conventional tagsets tend to ignore a broader range of contextual information such as "case" in the noun category. An information retrieval system, for instance, would benefit greatly from the contextual information embedded within a tag. When extracting information from a text, it is important to know whether the preposition preceding a noun indicates a "time initiation" or a "time termination" context. Other subtasks such as anaphora resolution may also benefit from contextual information within a noun category.

In PPT, however, any part of the tag may be used in the part of speech tagging project, while the rest of features may be marked as unknown. By doing so, every tagging project may be designed based on a different granularity level. In other words, it is not necessary for any tagging initiative to ensure that all feature positions are captured within a tag. This mechanism allows PPT tags to be easily converted to other tagging schemes. Similarly, any tagging project that uses PPT may be able to use existing annotation projects as its starting point and focus on refining the unknown feature values in the tag vector.

We believe that a standard positional tagset will enrich the information retrieval tasks by ensuring that various layers of information are embedded within the annotation tags.

References

Amtrup, J. W., Megerdoomian, K., & Zajac, R. (2000). Rapid Development of Translation Tools: Applications to Persian and Turkish. In *Proceedings of COLING 2000*. Saarbrucken, Germany.

Assi, M., & Hajiabdolhosseini, M. (2000). Grammatical tagging of a Persian corpus. *International Journal of Corpus Linguistics*, *5*(1), 69–81.

Assi, S. M. (2005). PLDB: Persian linguistics database. *Pa{ž}uhe{š}garan (Researchers)*.

Assi, S. M. (1997). FLDB: Farsi Linguistic Database. *International Journal of Lexicography*, *10*(3), 5.

Atwell, E. (2008). Corpus Linguistics: An International Handbook. In A. Lüdeling & M. Kytö (Eds.) (Vol. 29, p. Section IV, Chapter 23, pp.1–26). Mouton de Gruyter.

Bijankhan, M., javad Sheikhzadeh, Bahrani, M., & Ghayoomi, M. (2010). Lessons from Building a Persian Written corpus: Peykareh. *Language Resources and Evaluation*, *45*, 143–164. http://doi.org/10.1007/s10579-010-9132-x

Bijankhan, M., & Sheikhzadegan, J. (1994). FARSDAT-The Farsi Spoken Language Database. In *Proc. of the 5th Int. Conference of Speech Science and Technology* (Vol. 2, pp. 826–831). Perth, Australia.

Burnard, L. (Ed.). (2000). The British National Corpus Users Reference Guide, (May), 1–524. Retrieved from: http://www.natcorp.ox.ac.uk/

Dadegan Research Group (SCICT). (2012). *Persian Dependency Treebank Version 0.1 Annotation Manual and User Guide*.

Darrudi, E., & Hejazi, M. R. (2004). Assessment of a Modern Farsi Corpus. *The 2nd Workshop on Information Technology & Its Disciplines*.

Derzhanski, I. A., & Kotsyba, N. (2013). *Towards a consistent morphological tagset for Slavic languages: Extending MULTEXT-East for Polish, Ukrainian and Belarusian*.

Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.-. J., ı. Petkeviˇc, V., & Tufis, D. (1998).

Multext-East: Parallel and Comparable Corpora and Lexicons for Six Central and Eastern European Languages. In *Coling-acl* (pp. 315–319). Montr´eal, Qu´ebec, Canada.

Eeg-Olofsson, M. (1991). *Word-class tagging: Some computational tools*. University of Lund, Sweden.

Erjavec, T. (2012). MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language Resources and Evaluation*, *46*(1), 131–142.

Ghayoomi, M. (2004, July). *Word Prediction in Computational Processing of the Persian Language*. Islamic Azad University, Tehran Central Branch, Tehran, Iran. [In Persian]

Ghayoomi, M., Momtazi, S., & Bijankhan, M. (2010). A Study of Corpus Development for Persian. *International Journal on Asian Language Processing*, *20*(1), 17–33.

Greenbaum, S. (1992). The ICE tagset manual. In *Survey of English Usage*. London, UK.

Greenbaum, S. (1993). The Tagset for the International Corpus of English. *Corpus-Based Computational Linguistics*, *9*(9), 11–24.

Greene, B. B., & Rubin, G. M. (1971). *Automatic Grammatical Tagging of English*. Providence, R.I.

Hajic, J. (2004). Disambiguation of Rich Inflection: COMPUTATIONAL MORPHOLOGY OF CZECH (pp. 32–88). The University of Chicago Press.

Hana, J., & Feldman, A. (2010). A New Positional Tagset System for Russian. In *The 7th International Conference on Language Resources and Evaluation (LREC 2010)*.

Hardie, A. (2003). Developing a Tagset for Automated Part-of-speech Tagging in Urdu. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 Conference* (pp. 298–307).

Huckvale, M. (2004, July). SCRIBE - Spoken Corpus Recordings In British English. London, UK. Retrieved from: http://www.phon.ucl.ac.uk/resource/scribe/scribe-manual.htm

Johansson, S. (1986). The tagged LOB Corpus: User's Manual. Bergen, Norway: ICAME, The Computing Centre for the Humanities.

Khoja, S., Garside, R., & Knowles, G. (2001). A Tagset for the Morphosyntactic Tagging of Arabic. In P. Rayson, A. Wilson, T. McEnery, A. Hardie, & S. Khoja (Eds.), *Proceedings of the Corpus Linguistics 2001 Conference, UCREL Technical Paper 13* (p. 341).

Leech, G., & Wilson, A. (1996). Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Report EAG-TCWG-MAC/R. Retrieved from: http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html

Leech, G., & Wilson, A. (1999). Syntactic WordcWord Tagging. In H. van Halteren (Ed.) (pp. 55–80). KLUWER Academic Publishers.

QasemiZadeh, B., & Rahimi, S. (2006). Persian in MULTEXT-East Framework. In *FinTAL 2006: 5Th International Conference on Natural Language Processing* (pp. 541–551). Turku, Finland.

Rasooli, M. S., Moloodi, A., Kouhestani, M., & Minaei-Bidgoli, B. (2011). A Syntactic Valency Lexicon for Persian Verbs: The First Steps towards Persian Dependency Treebank. In *5th Language & Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics* (pp. 227–231).

Santorini, B. (1990). *Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision)*. University of Pennsylvania 3rd Revision 2nd Printing (Vol. 53). http://doi.org/10.1017/CBO9781107415324.004

Sawalha, M., & Atwell, E. (2013). A standard tag set expounding traditional morphological features for Arabic language part-of-speech tagging. *Word Structure*, *6*(1), 43–99. http://doi.org/10.3366/word.2013.0035

Sharifi-Atashgah, M., & Bijankhan, M. (2009). Corpus-based analysis for multi-token units in persian. *International Journal of Information and Communication Technology (Ijict)*, 1(3), 15–26.

Souter, C. (1989). A Short Handbook to the Polytechnic of Wales Corpus. Bergen, Norway: ICAME, The Norwegian Computing Centre for the Humanities.

Taylor, L. J., & Knowles, G. (1988). Manual of information to accompany the SEC corpus: The machine readable corpus of spoken English. Lancaster, UK: Unit for Computer Research on the English Language.