بسم الله الرحمن الرحیم

Persian Dependency Treebank
Version 0.1
Annotation Manual and User Guide

Dadegan Research Group

January 29, 2012

Contents

Chapter 1

Persian Dependency Treebank

1.1 Preface

Persian is a language with about 100 million speakers all over the world, yet in terms of the availability of teaching materials and annotated data for text processing, it is undoubtedly an under-resourced language. The need for more language teaching materials together with an ever-increasing need for Persian-language data processing have been the incentive for the inception of our project which has defined the development of the first ever syntactic treebank of Persian as its ultimate aim. A major by-product of the project has been the Persian verb valency lexicon [2] available free of charge for noncommercial uses. The present version of the corpus is a preliminary trial one aimed at introducing the project and attracting useful comments from users interested in the field.

In this manual, a brief introduction to the annotation schema of the treebank; i.e. dependency relations between Persian words, part of speech tags and morphosyntactic features, is presented.

1.2 Persian Dependency Treebank

This treebank is supplied for free noncommercial use. For commercial uses feel free to contact us. This is only a pre-version of the treebank which will be revised in the future. The number of annotated sentences will be increased to 30,000 sentences including samples from almost all verbs of the Persian valency lexicon.

1.2.1 Data Format

The data follows the format of CoNLL Shared Task on Dependency Parsing [1]. The morphosyntactic features include number, person, Tense/Mood/Aspect (for verbs) and word attachment status (1.4.2). To enable users to report bugs, a non-linguistic feature has also been added to the morphosyntactic ones: sentence id aligned with the treebank local database.

Abbreviation	Description
ACC. CASE MARKER	Accusative case marker
ENC. PR.	Enclitic pronoun
EZAFE	Ezafe marker
PAST	Past tense
PLUR	Plural
PRES	Present tense
SING	Singular
SUBJ	Subjunctive

Table 1.1: Description of abbreviations used in the manual.

1.2.2 How to Report Bugs

One of the features used alongside morphosyntactic features is senID. When you face an error in the annotation, please indicate the *senID* in the bug reporting page.

1.3 Dependency Relations

This section provides a brief introduction to the dependency relations. Table 1.1 shows the descriptions of the abbreviations used in this section. In Table 1.2 all dependency relations are listed.

1.3.1 Verb Dependents

SBJ: Subject

If there is an overt subject in a sentence, its relation with the head verb of the sentence is *SBJ*.

mæn	ketab	xandæm	**SBJ (xandæm, mæn)**[1]
I	book	read-PAST-1st-SING	

Translation: I read a book.

OBJ: Object

The object of Persian sentences may be identified by an accusative case marker "*ra*" which follows it. It is also possible for the object not to take it. The relation between the head verb and the object noun/pronoun (when "*ra*" is absent) or the accusative case marker is called *OBJ*.

[1] Dep(X, Y) means X is the head of Y with the relation of Dep.

mæn	ketab	xandæm		**OBJ (xandæm, ketab)**
I	book	read-PAST-1st-SING		

Translation: I read a book.

mæn	ketab	ra	xandæm	**OBJ (xandæm, ra)**
I	book	ACC. CASE MARKER	read-PAST-1st-SING	

Translation: I read a book.

NVE: Non-Verbal Element

Many Persian verbs are *compound verbs* (*complex predicates*). They are composed of at least two parts: one verbal and one non-verbal element. The non-verbal is a word belonging to noun, adjective, etc. class that conveys most of the lexical meaning of the compound verb. The relation between the verbal element and the non-verbal element is *NVE* in which the verbal element is the head.

ba	to	sohbæt	kærdæm	**NVE (kærdæm, sohbæt)**
with	you	speaking	do-PAST-1st-SING	

Translation: I spoke with you.

ENC: Enclitic Non-Verbal Element

In a number of Persian compound verbs, an enclitic pronoun which in person and number agrees with the subject appears after the non-verbal element. The relation between the verbal element and the non-verbal element of such compound verbs is called *ENC*. It should also be noted that the verbal element is always 3rd person singular.

æz	gæza	xoʃ	-æm	amæd	**ENC (amæd, xoʃ-æm)**
from	meal	good	1st-SING ENC. PR.	come-PAST-3rd-SING	

Translation: I liked the meal.

VPP: Prepositional Complement of Verb

Indirect object of verbs appears after a preposition. The relation between the verb and the complement preposition is *VPP*.

mæn	be	mædrese	ræftæm	**VPP (ræftæm, be)**
I	to	school	go-PAST-1st-SING	

Translation: I went to school.

OBJ2: Second Object

Second objects appear in sentences that seem to have two nominals as complements of their verbs. In such sentences, the noun that can potentially take a "*ra*" is the OBJ and one which can never have it is the *OBJ2*.

ketab	ra	be	ʔæli	hedje	dadæm	**OBJ2 (dadæm, hedje)**
book	ACC. CASE MARKER	to	Ali	gift	give-PAST-1st-SING	

Translation: I presented Ali with the book.

ketab	-i	be	ʔæli	hedje	dadæm	**OBJ2 (dadæm, hedje)**
book	INDEFINITE MORPHEME	to	Ali	gift	give-PAST-1st-SING	

Translation: I presented Ali with the book.

TAM: Tamiz

Tamiz is a property of an adjective or a noun ascribed to the subject (when object is absent) or to the object by the subject of a sentence whose main verb is some verbs like namidæn (= to name), xandæn (= to call), danestæn (= to consider), etc. The relation between the verb and *tamiz* is *TAM*.

ʔæli	ra	mærd	
Ali	ACC. CASE MARKER	man	
			TAM (mipendarim, xub)
-i	xub	mipendarim	
INDEFINITE MORPHEME	good	consider-PRES-1st-PLUR	

Translation: We consider Ali a good man.

MOS: Mosnad

Mosnad is a property of a noun, an adjective or a pronoun ascribed to the subject of a sentence whose main verb is a linking verb such as ʃodæn (= to become), budæn (= to be), ʔæstæn (= to be), etc. The relation between the verb and *mosnad* is *MOS*.

ʔu	doktor	ʔæst	**MOS (ʔæst, doktor)**
he	doctor	be-PRES-3rd-SING	

Translation: He is a doctor.

PROG: Progressive Auxiliary

Indicative present progressive and indicative preterite progressive tense-aspect-mood combinations in Persian are composed of two elements: the auxiliary (which is an inflected verb form of the infinitive"*daʃtæn*" agreeing with the main verb in person, number and tense) and the main verb. We posit that the auxiliary in such verbs is the dependent of the main verb. The relation is called *PROG*.

daʃtæm	miræftæm	**PROG (miræftæm, daʃtæm)**
have-PAST-1st-SING	go-PAST-PROG-1st-SING	

Translation: I was going.

ADVC: Adverbial Complement of Verb

Sometimes a noun referring to a time, a place, etc. may be the complement of a verb. The relation between the verb and the noun is *ADVC*.

tehran	mandæm	**ADVC (mandæm, tehran)**
Tehran	stay-PAST-1st-SING	

Translation: I stayed in Tehran.

VCL: Complement Clause of Verb

Some Persian verbs take clausal complements. The relation between such verbs and the head of the complement clause is called *VCL*. The head of the complement clause is usually a subordinating conjunction, but it may also be omitted in which case the head will be the main verb of the subordinate clause.

midanæm	ke	miʔajæd	**VCL (midænæm, ke)**
know- PRES-1st-SING	that	come-PRES-3rd-SING	

Translation: I know that he comes.

midænæm	miʔajæd	**VCL (midænæm, miʔajæd)**
know- PRES-1st-SING	come-PRES-3rd-SING	

Translation: I know he comes.

VPRT: Verb Particle

Some compound verbs in Persian have more than two parts, one part being the verbal element, the other a preposition and the last part a noun as the complement of the preposition. The relation between the verbal element and the preposition is called *VPRT*.

ɢodræt	be	dæst	ʔaværd	**VPRT (ʔaværd, be)**
power	to	hand	bring-PAST-3rd-SING	

Translation: He gained power.

LVP: Light Verb Particle

Verb forms derived from the compound infinitive "*pejda kærdæn*" (*to find*) may be used as a two-word light verb in some Persian compound verbs. In such cases, the relation between the verb forms of "*kærdæn*" and "*pejda*" is called *LVP*.

karxane	be	tehran	enteɢal	pejda	kærd	**LVP (kærd, pejda)**
factory	to	Tehran	transfer	visible	do-PAST-3rd-SING	

Translation: The factory was transferred to Tehran.

PARCL: Participle Clause

In coordination of two sentences with the same subject and different verbs of the same tense-aspect-mood, the first verb can be changed into past participle form. In such a case, we posit that the transformed verb is the dependent of the verb with normal inflection. The relation between the two is *PARCL*.

be	xane	ræfte	xabidæm
to	home	go-PAST ROOT+-e	sleep-PAST-1st-SING

PARCL (xabidæm, ræfte)

Translation: I went home and slept.

ADV: Adverb

As dependents of verb, adverbs specify the mode of action of the verb. Adverbs may be nouns, prepositions, adjectives functioning as adverbs, etc. the relation between the verb and the adverb is *ADV*.

bæraje	xærid	ræftæm
for	shopping	go-PAST-1st-SING

ADV (ræftæm, bæraje)

Translation: I went for shopping.

ʔæmdæn	ʃiʃe	ʃekæstæm	
intentionally	glass	break-PAST-1st-SING	**ADV (ʃekæstæm, ʔæmdæn)**

Translation: I broke the glass intentionally.

AJUCL: Adjunct Clause

Heads of all subordinate clauses enter a dependency relation with the verb of the main clause. In such cases, the relation between the verb in the main clause and the head of its dependent clause is *AJUCL*. The head of the adjunct clause is usually a subordinating conjunction, but it may also be omitted in which case the head will be the main verb of the subordinate clause.

ʔægær	bijaji	xoʃhal	miʃævæm	
if	come-SUBJ-2nd-SING	happy	become-PRES-1st-SING	**AJUCL (miʃævæm, ʔægær)**

Translation: If you come, I will become happy.

bijaji	xoʃhal	miʃævæm	
come-PRES-SUBJ-2nd-SING	happy	become-PRES-1st-SING	**AJUCL (miʃævæm, bijaji)**

Translation: If you come, I will become happy.

PART: Interrogative Particle

The words "*ʔaja*" and "*mægær*" are void of lexical meaning but can turn the sentence into a yes/no question. The relation between the main verb and the interrogative par-

ticle is called *PART*.

ʔaja	miʃenævi	**PART (miʃnævi, ʔaja)**
INTERROGATIVE PARTICLE	hear-PRES-2nd-SING	

Translation: Do you hear?

VCONJ: Conjunction of Verb

In sentence conjunctions, the main verbs of the sentences are coordinated. By convention we posit that the verb that appears last is the head of all others. The relation between a verb and a coordinating conjunction before it, is *VCONJ*. Conjunction of verb may also be established between two verbs if the coordinating conjunction is absent.

ræftæm	væ	xabidæm	**VCONJ (xabidæm, væ)**
go-PAST-1st-SING	and	sleep-PAST-1st-SING	

Translation: I went and slept.

1.3.2 Noun Dependents

NPREMOD: Pre-Modifier of Noun

Adjectives in their superlative form, pre-modifiers, pre-noun numerals and titles precede nouns and are considered pre-modifiers of the noun. The relation between a noun and its pre-modifier is *NPREMOD*.

behtærin	dust	**NPREMOD (dust, behtærin)**
best	friend	

Translation: the best friend.

ʔin	ketab	**NPREMOD (ketab, ʔin)**
this	book	

Translation: this book.

NPOSTMOD: Post-Modifier of Noun

Adjectives in their positive and comparative forms together with post-noun numerals are considered post-modifiers of noun. The relation between a noun and its post-modifier is *NPOSTMOD*.

ketab	-e	xub	**NPOSTMOD (ketab, xub)**
book	EZAFE	good	

Translation: the good book.

NPP: Preposition of Noun

Regardless of whether the preposition is an adjunct or a complement, its relation with the head noun is called *NPP*.

jedal	dar	tasuki	**NPP (battle, in)**
battle	in	Tasooki	

Translation: battle in Tasooki.

ʔetteka	be	valedejn	**NPP (ʔetteka, be)**
dependence	to	parents	

Translation: dependence on parents.

NCL: Clause of Noun

Clauses which function as dependents of nominal heads can be either their complements or their adjuncts. The relation between a noun and both types of clausal dependents is *NCL*.

mærd	-i	ke	didi	**NCL (mærd-i, ke)**
man	INDEFINITE MORPHEME	that	see-PAST-2ND-SING	

Translation: the man you saw.

MOZ: Ezafe Dependent

Ezafe dependents in Persian are nouns or pronouns which follow a head noun and signify a posses-possessor, first name-last name, etc. relation with the head noun. The sign for an ezafe construction in Persian is a vowel /e/ which is pronounced right after the head noun, but is usually absent in the Perso-Arabic script. The relation between a noun and its ezafe dependent is *MOZ*.

ketab	-e	hæsæn	**MOZ (ketab, hæsæn)**
book	EZAFE	Hasan	

Translation: Hassan's book.

APP: Apposition

An apposition is a noun which follows another noun or a pronoun and has the same reference as the first and they both have the same syntactic function. When a noun comes in apposition with another noun or pronoun, the first is considered the head and the second, the dependent.

sæʔdi	ʃaʔer	-e	ʔirani	**APP (sæʔdi, ʃaʔer-e)**
Saadi	poet	EZAFE	Iranian	

Translation: Saadi, the Iranian poet.

NCONJ: Conjunction of Noun

When two nouns become related by a coordinating conjunction, a relation is established between the first (=head) noun and the coordinating conjunction. This relation

is *N*CONJ. Conjunction of noun may also be established between two nouns if the coordinating conjunction is absent.

sæʔdi	væ	hafez	**NCONJ (sæʔdi, væ)**
Saadi	and	Hafez	

Translation: Saadi and Hafez.

NADV: Adverb of Noun

The relation between a noun and a modifying adverb: When the verb is a complex predicate, in some cases, the adverbial concept is expressed without using a preposition and the meaning of a preposition plus a complement is understood. In such a case, we assume that the complement of the omitted preposition comes into a relation with the non-verbal element of the complex predicate (whether we should draw a dependency arc between the complement and the non-verbal element or between the complement and the verbal element, depends on the semantics of the sentence).

tehran	sokunæt	daræm	**NADV (sokunæt, tehran)**
Tehran	residence	have-PRES-1st-SING	

Translation: I reside in Tehran.

NE: Non-Verbal Element of Infinitive

It is possible for all Persian verbs to be transformed to their corresponding infinitives. Infinitives in Persian show the syntactic behavior of nouns. Given the fact, a complex predicate which is transformed to its corresponding infinitive, retains the relationship between the non-verbal element and the infinitival form of the verbal element. The relation between a noun transformed from a complex verb and its non-verbal element is *N*E.

ʔexradʒ	kærdæn	**NE (kærdæn, ʔexradʒ)**
sacking	to do	

Translation: to fire.

MESU: Measure

In some cases, nouns (countable and uncountable) are preceded by another noun which serves as a counting unit. The counting unit itself might be preceded by a pre-noun numeral or followed by an indefinite morpheme /-i/.

do	dʒeld	ketab	**MESU (ketab, dʒeld)**
two	volume	book	

Translation: two volumes of book (two books).

NPRT: Particle of Infinitive

As explained earlier, all Persian verbs can be converted to their corresponding infinitives and be used in sentences where they function as nouns. Some verbs in Persian contain prepositions as their integral parts. After being converted to infinitives, they retain their prepositional elements. The dependency relation between an infinitive and its prepositional element is *NPRT*.

æz	dæst	dadæn	**NPRT (dadæn, æz)**
from	hand	to give	

Translation: to lose.

1.3.3 Adjective Dependents

COMPPP: Comparative Preposition

Comparative forms of adjectives and adverbs in Persian need the preposition "*æz*" to introduce the second member of an unequal comparison. The relation between the comparative adjective or adverb and "*æz*" is called *COMPPP*.

behtær	æz	servæt	**COMPPP (behtær, æz)**
better	than	welath	

Translation: better than wealth.

ADJADV: Adverbial Complement of Adjective

In cases where the complement preposition of an adjective is omitted, the relation between the adjective and the complement of the deleted preposition is called *A*JADV.

taksi	sævar	ʃodæm	**ADJADV (sævar, taksi)**
taxi	riding	become-PAST-1st-SING	

Translation: I got in a taxi.

ACL: Complement Clause of Adjective

Adjectives may have clausal complements. The relation between the adjective and the head of the clause is called *A*CL.

ʔagah	hæstæm	ke	miʔaji	**ACL (ʔagah, ke)**
aware	be-PRES-1st-SING	that	come-PRES-3rd-SING	

Translation: I am aware that you will come.

AJPP: Prepositional Complement of Adjective

Adjectives may have prepositional complements. The relation between the adjective and the preposition is called *A*JPP.

ʔaʃna	ba	ʔækkasi	**AJPP (ʔaʃna, ba)**
familiar	with	photography	

Translation: familiar with photography.

NEZ: Ezafe Complement of Adjective

Adjectives and their nominal complements may enter an Ezafe construction in which the adjective is the head and the noun is the dependent. In Persian a vowel /e/ is pronounced right after the adjective.

negæran	-e	ʔu	**NEZ (negæran-e, ʔu)**
anxious	EZAFEH	him	

Translation: anxious about him.

AJCONJ: Conjunction of Adjective

The relation between an adjective and a coordinating conjunction is called *AJCONJ*. Conjunction of adjective may also be established between two adjectives if the coordinating conjunction is absent.

ʃad	væ	særzende	**AJCONJ (ʃad, væ)**
happy	and	lively	

Translation: happy and lively.

APREMOD: Adjective Pre-Modifier

Adjectives may be modified by adverbs. In such cases, the relation between the adjective and the modifying adverb is called *APREMOD*.

besjar	ʃad	**APREMOD (ʃad, besjar)**
very	happy	

Translation: very happy.

APOSTMOD: Adjective Post-Modifier

Adjectives may be modified by adjectives. In such cases, the relation between the modified adjective and the modifier is called *APOSTMOD*.

pirahæn	-e	ʔabi	-je	ʔasemani	**APOSTMOD (ʔabi-je, ʔasemani)**
shirt	EZAFE	blue	EZAFE	skiey	

Translation: a sky blue shirt.

1.3.4 Other Dependents

PREDEP: Pre-Dependent

Here we introduce some of the commonest uses of this dependency label but it has to be noted that in cases where defining a unique dependency label does not seem economical, we use *PREDEP* for all dependents that come before their heads.

The most common pre-dependent is the relation between the only Persian postposition "*ra*" and the direct object of the verb. In most cases in Persian sentences where there is a direct object the free morpheme "*ra*" follows the direct object. In the present corpus we posit that "*ra*" is the head of the direct object.

æli	ra	didæm	**PREDEP (ra, æli)**
Ali	ACC. CASE MARKER	see-PAST-1ST-SING	

Translation: I saw Ali.

Another common use of *PREDEP* refers to the relation between a coordinating conjunction and the verb preceding it. By convention we posit that in sentences bearing verb coordination the last verb is the head.

xandæm	væ	neveʃtæm	**PREDEP (væ, xandæm)**
read-PAST-1ST-SING	and	write-PAST-1ST-SING	

Translation: I read and wrote.

In cases of infinitives used as nouns in Persian sentences, we consider all their preceding dependents, other than NPRT's and NE's to be their pre-dependents.

ʃad	kærdæn	**PREDEP (kærdæn, ʃad)**
happy	to do	

Translation: to make happy.

There are words like "*hætta*", "*hæm*", and "*næ*" which modify the words they precede. If their following words belong to lexical classes other than verb, they are considered to be their PREDEP's.

hætta	ʔæli	fæhmid	**PREDEP (ʔæli, hætta)**
even	Ali	learn-PAST-3rd-SING	

Translation: Even Ali learnt.

POSDEP: Post-Dependent

Again, if defining a new dependency label is not economical we use *POSDEP* for all dependents that come after their heads. We introduce two of the commonest uses of the label.

Objects of all prepositions are post-dependents.

ketab	ra	be	ʔæli	dadæm	**POSDEP (be, ʔæli)**
book	ACC CASE-MARKER	to	ali	give-PAST-1st-SING	

Translation: I gave the book to Ali.

Another common use of *POSDEP* refers to the relation between a coordinating conjunction and coordinated word after it (it may be a noun, an adjective an adverb, a preposition but not a verb).

xub	væ	mofid	**POSDEP (væ, mofid)**
good	and	useful	

Translation: good and useful .

PCONJ: Conjunction of Preposition

Two or more prepositions may be coordinated using coordinating conjunctions. The relation between a preposition and a coordinating conjunction following it is called *PCONJ*.

dær	tehran	væ	ba	ma	bud	**POSDEP (dær, væ)**
in	Tehran	and	with	us	be-PAST-3rd-SING	

Translation: He was in Tehran and with us.

AVCONJ: Conjunction of Adverb

The relation between an adverb and a coordinating conjunction in sentences where two or more adverbs are conjoined is called *AVCONJ*. Conjunction of adverb may also be established between two adverbs if the coordinating conjunction is absent.

maherane	væ	ziba	minevisæd	**AVCONJ (maherane, væ)**
skillful	and	beautiful	write-PRES-3rd-SING	

Translation: He writes skillfully and beautifully.

PRD: Predicate

The head in subordinate or relative clauses is the subordinating conjunction and its relation with the main verb of the clause is *PRD*.

ʔamædæm	ta	bebinæm	**PRD (ta, bebinæm)**
come-PAST-1st-SING	to	see-PRES-1st-SING	

Translation: I came to see.

ROOT: Sentence Root

The head of the whole sentence (usually a verb) is itself headed by an abstract element. This relation is called *ROOT*.

PUNC: Punctuation Mark

Full stops indicating the end of a sentence are dependents of the head of the whole sentence. Punctuation marks like ",", ";", ":", etc. are dependents of words immediately preceding them. Punctuation marks that appear in pairs like "()" are dependents of the head word within them.

1.4 Part of Speech Tags and Morphosyntactic Features

Parts of Speech (henceforth POS's) are classifications of words based on their functions in sentences for purposes of grammatical analysis. In this manual, 17 [coarse-grained] POS's have been recognized for Persian words. Each [coarse-grained] POS is divided into a number of fine-grained POS's. In cases where no fine-grained POS has been recognized, the fine-grained POS is the same as the coarse grained one. Moreover, there are a number of properties that might be active for each POS and if active, they have a number of values. Table 1.3 shows how the above-mentioned classification works.

1.4.1 Coarse-grained and Fine-grained POS tags

ADJ: Adjective

An adjective is a word that modifies or qualifies a noun.

- **AJP (Positive)**: The positive form of an adjective is used when the adjective is not meant to accomplish any comparison.
- **AJCM (Comparative)**: The comparative form of an adjective is used for a comparison between two entities.
- **AJSUP (Superlative)**: The superlative form of an adjective is used for a comparison among more than two entities.

ADR: Address Term

Morphemes that accompany a noun to make it the address of the speaker are called address terms.

- **PRADR (Pre-noun)**: Pre-noun address terms precede nouns as single words.
- **Post-noun (POSADR)**: Post-noun address terms follow nouns as bound .

ADV: Adverb

Adverbs typically specify the mode of action of the verb.

- **SADV (Genuine)** : Genuine adverbs are single word-forms that can only function as adverbs.
- **AVP (Positive)**: Positive adverbs are positive forms of adjectives that modify verbs rather than nouns.
- **AVCM (Comparative)**: Comparative adverbs are Comparative forms of adjectives that modify verbs rather than nouns.

CONJ: Coordinating Conjunction

Coordinating conjunctions are a class of words that connect words. Words connected via coordination are usually of equivalent syntactic status.

IDEN: Title

Titles are respectful words used together with people's first or last names. Depending on what title is being used, they may precede or follow the name.

N: Noun

A noun is usually defined as a word denoting a thing, place, person, quality, or action and functioning in a sentence as the subject or object of a verb or as the object of a preposition.

- **ANM (Animate)**: A class of nouns whose reference is to persons, animals and plants.
- **IANM (Inanimate)**: A class of nouns whose reference is to anything other persons, animals and plants.

PART: Particle

Persian particles reflect the mood or attitude of the speaker and highlight the sentence focus.

POSNUM: Post-noun Numeral

Ordinal numbers ending in *-om* follow the nouns they modify.

POSTP: Postposition

Postpositions are a class of words that are used after nouns or pronouns. In Persian the only postposition is "*ra*".

PR: Pronoun

Pronouns are a class of words that refer to the closed set of items which can be used to substitute for nouns.

- **SEPER (Separate Personal)**: Separate personal pronouns are a subclass of pronouns that are separate (not connected to other words orthographically) and personal (refer to 1st, 2nd or 3rd persons).
- **JOPER (Enclitic Personal)**: Enclitic personal pronouns are personal pronouns that are connected to the end of a verb and function as its object.
- **DEMON (Demonstrative)**: A subclass of pronouns that point to an entity in the situation or elsewhere in a sentence is called demonstrative pronoun.
- **INTG (Interrogative)**: An interrogative pronoun is a pronoun used in order to ask a question. Some of them refer to people others refer to people and objects, etc.
- **CREFX (Common Reflexive)**: Three reflexive pronouns are used in Persian: "*xod*", "*xiʃ*" and "*xiʃtæn*", all meaning "*self*". The three forms are used for all persons and numbers.
- **UCREFX (Noncommon Reflexive)**: Another type of reflexive pronoun is UCREFX that inflects for person and number. There are six UCREFX's in Persian.
- **RECPR (Reciprocal)**: Reciprocal pronouns express the meaning of mutual relationship.

PREM: Pre-modifier

Pre-modifiers are a class of noun modifiers that precede nouns and are in complementary distribution with other members of the class.

- **EXAJ (Exclamatory)**: Exclamatory pre-modifiers express the speaker's surprise or other emotional attitudes toward the noun being modified.
- **QUAJ (Interrogative)**: Interrogative pre-modifiers question the noun being modified.
- **DEMAJ (Demonstrative)**: Demonstrative pre-modifiers point to the noun being modified as something or someone in proximity or in distance.
- **Ambiguous (AMBAJ)**: Ambiguous pre-modifiers modify their following nouns without specification or identification.

PRENUM: Pre-noun Numeral

In Persian, cardinal numbers and ordinal numbers ending in *-omin* suffix precede the nouns they modify.

PREP: Preposition

Prepositions are a class of words that are used before nouns or pronouns functioning as modifiers of verbs, nouns, or adjectives, and that typically express a spatial, temporal, or other relationships.

PSUS: Pseudo-Sentence

Persian pseudo-sentences are a class of words that when used, do not let the sentence have a verb. In fact, they compensate for the lack of verb.

PUNC: Punctuation Mark

All punctuation marks are assigned the *PUNC* part of speech tag.

V: Verb

Defined formally, a verb is an element capable of showing morphological contrasts of tense, aspect, voice, mood, person and number. The verbs of Persian have been subdivided into three fine grained parts of speech.

- **ACT (Active)**: The form of the verb in which the subject is the actor of the verb is called the active voice form of the verb. Active verbs constitute the majority of verbs in Persian.
- **PASS (Passive)**: The form of the verb in which the subject is the undergoer of the action of the verb and where an inflected form of the auxiliary "ʃodæn" is present is known as the passive voice form of the verb.
- **MOD (Modal)**: Modals in Persian are a class of verbs which denote notions of uncertainty, definiteness, vagueness, possibility, etc. The most important grammatical behavior of Persian modals is that they do not inflect for all persons and/or tense-aspect-mood combinations.

SUBR: Subordinating Conjunction

Subordinating conjunctions are a class of words that connect words. Words connected via subordination are of different syntactic status.

1.4.2 What is Word Attachment?

In some special cases, we are obliged to break a word into smaller parts in order to capture the syntactic relations between the sentence elements. For example, the word *didæmæʃ* should be broken into two parts: 1) *didæm*, and 2) *æʃ*, because *æʃ* plays the role of the object of the verb *didæm*. As another example, the orthographic word *mæra* should be broken into two parts: 1) *mæ* (contracted form of the personal pronoun *mæn*), and 2) *ra*, because *ra* is a postposition in the sentence and may play the role of the object or the complement adposition of the verb. Table 1.5 shows more details about word attachment.

Abbreviation	Description
ACL	Complement Clause of Adjective
ADV	Adverb
ADVC	Adverbial Complement of Verb
AJCONJ	Conjunction of Adjective
AJPP	Prepositional Complement of Adjective
AJUCL	Adjunct Clause
APOSTMOD	Adjective Post-Modifer
APP	Apposition
APREMOD	Adjective Pre-Modifier
AVCONJ	Conjunction of Adverb
COMPPP	Comparative Preposition
ENC	Enclitic Non-Verbal Element
LVP	Light Verb Particle
MESU	Measure
MOS	Mosnad
MOZ	Ezafe Dependent
NADV	Adverb of Noun
NCL	Clause of Noun
NCONJ	Conjunction of Noun
NE	Non-Verbal Element of Infinitive
NEZ	Ezafe Complement of Adjective
NPOSTMOD	Post-Modifer of Noun
NPP	Preposition of Noun
NPREMOD	Pre-Modifier of Noun
NPRT	Particle of Infinitive
NVE	Non-Verbal Element
OBJ	Object
OBJ2	Second Object
PARCL	Participle Clause
PART	Interrogative Particle
PCONJ	Conjunction of Preposition
POSDEP	Post-Dependent
PRD	Predicate
PREDEP	Pre-Dependent
PROG	Progressive Auxiliary
PUNC	Punctuation Mark
ROOT	Root
SBJ	Subject
TAM	Tamiz
VCL	Complement Clause of Verb
VCONJ	Conjunction of Verb
VPP	Prepositional Complement of Verb
VPRT	Verb Particle

Table 1.2: Dependency relations in the Persian dependency treebank

Morphosyntactic features in the Persian dependency treebank				
CPOS	**FPOS**	**Person**	**Number**	**TMA**
ADJ	AJP AJCM AJSUP			
ADR	PRADR POSADR			
ADV	SADV AVP ADCM			
CONJ				
IDEN				
N	ANM IANM		SING PLUR	
PART				
POSNUM				
POSTP				
PR	SEPER JOPER DEMON INTG CREFX UCREFX RECPR	1 2 3	SING PLUR	
PREM	EXAJ QUAJ DEMAJ AMBAJ			
PRENUM				
PREP				
PSUS				
PUNC				
V	ACT PAS MOD	1 2 3	SING PLUR	See Table 1.4
SUBR				

Table 1.3: Morphosyntactic features in the Persian dependency treebank. Empty cells indicate that the mentioned feature is not present for the POS. *TMA* stands for *Tense/Mood/Aspect*; *CPOS* for *Coarse grained POS* and *FPOS* for *Fine grained POS*.

Tense/Aspect/Mood	**Abbreviation**	**Examples** xordæn: to eat
Imperative	HA	boxor
Indicative Future	AY	xahæm xord
Indicative Imperfective perfect	GNES	mixordeæm
Indicative Imperfective pluperfect	GBES	mixorde budæm
Indicative Imperfective preterite	GES	mixordæm
Indicative Perfect	GN	xordeæm
Indicative Pluperfect	GB	xorde budæm
Indicative Present	H	mixoræm
Indicative Preterite	GS	xordæm
Subjunctive Imperfective pluperfect	GBESE	mixorde bude baʃæm
Subjunctive Imperfective preterite	GESEL	mixorde baʃæm
Subjunctive Pluperfect	GBEL	xorde bude baʃæm
Subjunctive Present	HEL	boxoræm
Subjunctive Preterite	GEL	xorde baʃæm

Table 1.4: Tense/Mood/Aspect types in Persian verbs

Attachment State	**Abbreviation**
Isolated word	ISO
Attached to the word on the right	PRV
Attached to the word on the left	NXT

Table 1.5: Attachment states in Persian words in the dependency treebank

Bibliography

[1] Sabine Buchholz and Erwin Marsi. CoNLL-X shared task on multilingual dependency parsing. In *Proceeding of the Tenth Conference on Computational Natural Language Learning (CoNLL)*, 2006.

[2] Mohammad Sadegh Rasooli, Amirsaeid Moloodi, Manouchehr Kouhestani, and Behrouz Minaei-Bidgoli. 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*, pages 227–231, Poznan, Poland.

Appendix A

Dadegan Research Group

A.1 About Dadegan Research Group

Dadegan research group is funded by Supreme Council of Information and Communication Technology (SCICT) in order to fulfill the needs of feasible annotated data for the Persian language. The main concern of this group is to promote Persian language in the computational processing environments.

Contact: info@dadegan.ir

A.1.1 Project Members

- **Project Head and Computational Linguistic Research**
  - Mohammad Sadegh Rasooli: MSc. Student, Iran University of Science and Technology
- **Linguistic Research and Instruction**
  - Manouchehr Kouhestani: MA, University of Tehran
  - Amirsaeid Moloodi: PhD candidate, University of Tehran
- **Linguistic Annotation**
  - Farzaneh Bakhtiary: MA student, University of Tehran
  - Parinaz Dadras: MA student, University of Tehran
  - Maryam Faal-Hamedanchi: PhD, Peoples' Friendship University of Russia
  - Saeedeh Ghadrdoost-Nakhchi: MA student, University of Tehran
  - Mostafa Mahdavi: PhD candidate, Institute for Humanities and Cultural Studies, Tehran
  - Azadeh Mirzaei: PhD candidate, Allameh Tabatabaei University, Tehran
  - Neda Poormorteza-Khameneh: MA, Islamic Azad University

  - Morteza Rezaei-Sharifabadi: MA student, Sharif University of Technology
  - Akram Shafie: MA, University of Tehran
  - Salimeh Zamani: MA, Islamic Azad University

- **Programming Support**
  - Seyed Mahdi Hoseini: MSc. student, Iran University of Science and Technology
  - Alireza Noorian: MSc. student, Iran University of Science and Technology
  - Yasser Souri: MSc. student, Sharif University of Technology

- **Web Support**
  - Mohsen Hossein-Alizadeh, SCICT

- **Persian Language and Orthography Committee Head, SCICT**
  - Mahdi Behniafar: PhD, Asistant Professor, Allameh Tabatabaei University, Tehran

A.2 Acknowledgment

We gratefully appreciate all the comments by the users of Dadegan products and hope that this interaction remains and increases. We also appreciate Faezeh Abbasi-Abyaneh, Seyedeh Maneli Hashemian, Shima Zamanpoor, Narmin Ghaderi and Hura Nuri for their useful help to find sample sentences of inflections of rare verbs in Persian.