Corpus-based Analysis for Multi-Token Units in Persian

Massoud Sharifi-Atashgah
Linguistics Department
The Faculty of Letters and Humanities
Tehran University
Enghelab St. Tehran, Iran
massoud.sharifi@ymail.com

Mahmood Bijankhan
Linguistics Department
The Faculty of Letters and Humanities
Tehran University
Enghelab St. Tehran, Iran
mbjkhan@ut.ac.ir

Abstract

Morphological and syntactic annotation of multi-token units confront several problems due to the concatenating nature of Persian script and so its orthographic variation. In the present paper, by the analysis of the different collocation types of the tokens, the compositional, non-compositional and semi-compositional constructions are described and then, in order to explain these constructions, the static and dynamic multi-token units will be introduced for the non-generative and generative structures of the verbs, infinitives, prepositions, conjunctions, adverbs, adjectives and nouns. Defining the multi-token unit templates for these categories is one of the important results of this research. The findings can be input to the Persian Treebank generator systems. Also, the machine translation systems using the rule-based methods to parse the texts can utilize the results in text segmentation and parsing.

1 Introduction

Since in Arabic script-based languages such as Persian the script is exclusively written cursively and the different orthographic forms exist for most of the letters, the multi-token units (from now on MTUs) are usually written in concatenated or non-concatenated ways. In the latter, a linguistic unit is considered as several words and consequently they get different part of speech (POS) tags in comparison with the concatenated case. This non-integrity in the POS tagging causes several problems in the generation of the Persian Treebank and the machine translation systems using the rule-based methods to parse the texts, as well.

Corpus-based identification and analysis for MTUs in Persian result in the creation of a *segmentation* subsystem in the Treebank and machine translation systems. In the present study, by utilizing the information existing in the POS-tagged texts of the Contemporary Persian Corpus, we analyze the Persian static and dynamic MTUs for most of the major categories and show that how these MTUs can affect the Persian Treebank and machine translator generation process.

The design and implementation of a system require the system inputs to be specified. Persian texts alongside their POS tags are considered as the input of a processing system. By using the segmentation and bracketing modules inside the system, the Persian Treebank can be generated as the output of the mentioned system.

The Penn Treebank with more than 4.5 million words was based on the pioneering Brown Corpus and became the basis of the following studies. The Penn Treebank aimed at the word-level and phrase-level annotation of the texts. For the reasons of recoverability and consistency, the Penn Treebank project team pared down the Brown Corpus tagset to 48 tags (Marcus et al., 1993).

A further difference between the Penn Treebank and the Brown Corpus concerns the significance accorded to syntactic context. That is, a word may get different POS tags in different syntactic contexts. The syntactic tagset of the Penn Treebank contains 14 tags including NP, ADJP, PP, VP, S and SBAR. In addition to the delineation of the phrasal boundaries in the Penn Treebank,

any phrase can receive some more functional tags including syntactic, semantic and topicalization tags.

One group of these functional tags is related to the adverbials being usually the VP adjuncts. For instance, the adverbial of manner (-MNR).

Based on the Penn English Treebank, other Treebanks have emerged including the Penn Arabic Treebank. Although the framework of these Treebanks is the Penn English Treebank, having considered the Arabic and Hebrew language-specific properties, some modifications have been fulfilled on the morphological and syntactic annotation methods. (Bies and Maamouri, 2003).

In Iran, the last corpus produced by the Research Center for Intelligent Signal Processing (RCISP), is the *Text Corpus of Contemporary Persian* (from now on, abbreviated to ***Peykare***) (Bijankhan et al., 2008).

The organization of the present paper is as follows. In section 3, we introduce Peykare in brief and review its major (obligatory) and peripheral POS tags. Persian script problems are addressed in section 3 first and then the static and dynamic MTUs are described. Static MTU forming a closed category is the subject of section 4 where we discuss the MTUs in the major categories like verbs, infinitives, prepositions, conjunctions, adverbs, adjectives and nouns. In section 5, the dynamic MTUs are discussed. Concluding remarks follow in the last section.

2 Introducing "Peykare" and Tagsets

110 million words of the contemporary Persian language have been randomly chosen, from different sources, and with various sizes. Moreover, about ten million words were randomly selected and POS-tagged in Peykare.

Different tagsets have been used in the corpora due to the language-specific considerations (Cloeren, 1999). Four issues were taken into account in designing the Peykare morpho-syntactic tagset: word definition, morphology, homographs and goals. For fulfilling the goals, EAGLES Guidelines were used. One of the important goals was paving the way for tagging the syntactic phrases and generating the Persian Treebank.

EAGLES Guidelines provide thirteen obligatory (major) categories, which are equivalent with the parts of speech in the traditional grammar. Also, for each major category, at the most, three peripheral attributes have been devised: Recommended, Generic and language-specific attributes (Leech and Wilson, 1996; Bijankhan et al., 2008).

The major categories of Peykare are: Noun (N), Verb (V), Adjective (ADJ), Pronoun (PRO), Determiner (DET), Adverb (ADV), Preposition (PREP), Postposition (POSTP), Conjunction (CONJ), Numeral (NUM), Interjection (INT), Residual (RES), Classifier (CL) and Punctuation (PUNC). The above-mentioned major categories are the essential building blocks of the phrases.

As mentioned before, one part of the peripheral categories is the recommended attributes. For instance, nouns have Common (COM), Proper (PR), Singular (SING) and Plural (PL) attributes and Pronouns own demonstrative (DEM) and indefinite (INDF) attributes. (For more tags refer to Bijankhan et. al, 2008).

3 Persian Script Problems and Multi-Token Units (MTUs)

In the present section we first show the Persian script problems which are related to the MTUs and then describe the MTU and its static and dynamic types.

3.1 Persian Script and Its Problems

After the Latin alphabet, Arabic alphabet is the second-most widely used one in the world. Non-Semitic languages like Persian, Pashto and Urdu are written with the Arabic alphabet. Persian alphabet that is also called Perso-Arabic script has five more letters than standard Arabic: پ [p], چ [ʧ], ژ [ʒ], and گ [g] (Lazard, 1992) and ء[ʔ] (Hamze). Thus, the size of Persian alphabet adds up to 33 letters. It is exclusively written cursively from right to left. That is, the majority of letters in a word connect to each other. In Persian script, 25 letters have four forms based on their positions in a chain of letters. They have *isolated*, *initial*, *medial* and *final* forms (Lazard, 1992). Hamze is also a *joining* letter. These 26 joining letters are: پ، ب، ء ق، ف، غ، ع ،ظ، ط، ض، ص، ش، س، خ، ح، چ، ج، ث، ت، ه، ن، م، ل، گ، ک، and ی. Other letters which are called *disjoint* letters are: ژ، ز، ر، ذ، د، ا and و.

Moreover, the Persian orthography allows some morphemes to appear as bound or free affixes

before or after a morpheme (Megerdoomian, 2000). Therefore, many words can be written as concatenated or non-concatenated. In the concatenated form, the initial or medial form of the first morpheme is joined to the medial or final form of the second morpheme. For instance, in the word کتابها [ketɑb-hɑ](books), the initial form بـ of the morpheme کتاب[ketɑb] (book), is joined to the medial form ـها of the plural morpheme ها[hɑ]. In the non-concatenated form, the space or Zero Width Non-Joiner (ZWNJ) character (shown as '|') is inserted and the final form of the first morpheme is joined to the initial form of the second one: کتاب ها and کتاب‌ها.

In the definition of *word* as a linguistic unit, there is not a consensus among the linguists, but it is possible to define the word in any language with an acceptable precision (Graff, 1929; Garvin, 1954; Chomsky and Halle, 1968).

In Persian texts, a word can be considered as a chain of letters which make up at least one free morpheme such that, regardless of its affixes and enclitics, last syllable bears stress. Typists intuitively and somehow unanimously can recognize words according to the above definition, like other literates. However, while typing texts, they do not separate words in the same manner even with following the Persian orthography which has been published by the Persian Academy of Language and Literature (PALL). The PALL supports the non-concatenated writing in Persian orthography if the ZWNJ character is inserted between building morphemes (Persian Academy of Language and Literature, 2005).

The above-mentioned points about the Perso-Arabic script and its several letter types are the major reasons for the emergence of the *orthographic variation* (Buckwalter, 2004).

Suppose that each *orthographic token* is specified by a *delimiter* that is normally a space character or a punctuation mark. Also presume that a token can be a morpheme, a simple word, an inflection, a derivation or a compound. Therefore, a many-to-many relationship is created between an orthographic token and a linguistic unit. That is, an orthographic token may show one or several linguistic units. For example, the token کزین, [kæzin] constitutes three linguistic units: a conjunction که, [ke], "that", a preposition از,[ʔæz], "from" and a pronoun این, [ʔin], "this". This case is called a Multi Unit Token (MUT). On the other hand, a linguistic unit may be written as one or several tokens and get several POS tags in Peykare wrongly. This case is an instance of an MTU.

3.2 MTUs in Persian

If some tokens make up a linguistic unit unitedly, then we will have an MTU (Cloeren, 1999). In fact, MTUs are words that should have one POS tag. In Peykare, if an MTU is written as concatenated or with ZWNJ character (the recommendation of the PALL), it usually receives just one POS tag, e.g., بنابراین, [bænɑbærʔin], "so" has CONJ tag, به‌سختی, [besæXti], "hard, hardly" has ADV tag. But if these tokens are written with space character, then each token gets a separate POS tag, e.g., بنا بر این is considered as three words and gets three tags: بنا ,[bænɑ], "basis", (POS tag: N, COM, SING), بر, [bær], "on", (POS tag: PREP) and این, [ʔin], "this", (PRO,DEM,SING) or به سختی is treated as two units and two tags: به, [be], "to", (POS tag: PREP) and سختی, [sæXti], "hardship", (POS tag: N,COM,SING). (see section 3.1 for the POS tags).

In Peykare, when some tokens are bound morphemes that cannot appear in the text alone and so get separate tags, e.g., plural morphemes(ها [hɑ], جات [dʒɑt]), imperfective prefix(می [mi]) or when the tokens constitute a collocation most of the times, e.g. دوچرخه سوار, [doʃærXe sævɑr], "biker", بالا پوش, [bɑlɑ puʃ], "overcoat", the tokens are correctly tagged as a single unit.

Before proceeding with the MTU problems, it is necessary to analyze the various types that the tokens may be collocated. Four types can be supposed for the collocated tokens.

Type 1: each of the tokens has the morphological and semantic functions and the tokens together build a phrase. The meaning of the phrase is the sum of the meanings of the tokens. This is called a *compositional construction*, e.g., در کتاب دستورزبان, [dær ketɑb-e dæsturzæbɑn], "in the grammar book", is a PP with compositional meaning.

Type 2: the tokens constitute a phrase, but the phrase has its own morphological and semantic functions and the phrase meaning is not the composition of the meanings of the tokens. This is called a non-compositional construction, e.g., the complex predicates such as داد زدن, [dɑd zædæn], (scream hitting), "to yell" and the adverb به مراتب,

[be mærɑteb], (to stages), "by far".

Type 3: the tokens build a phrase that has both a compositional and a non-compositional meaning. As mentioned above, [be sæXti] is a PP with two tokens. This PP in (1) has a compositional meaning but in (2) has a non-compositional meaning and a semantic tag of the adverbial of manner (-MNR).

(1) ساسان به سختی عادت کرد
[NP Sassan] [PP be sæXti][VP ʔɑdæt kærd]
Sassan got used to the hardship

(2) ساسان به سختی کار کرد
[NP Sassan][PP-MNR be sæXti][VP kɑr kærd]
Sassan worked hard.

Type 4: in this type, the tokens in a phrase can collocate with certain templates, but not completely freely, and create a compositional meaning that is similar to type 1 (compositional construction). However, at the same time, the phrase can also get a function tag and become like type 2 (non-compositional construction). We call this type *semi-compositional construction*. For instance, سه نفره [se næfære] (threesome, by/for three persons), ده کیلویی [dæh kiluʔi] (ten-kilo) which are phrases with adjectival function.

In analyzing these types, we saw that the first type builds an open and generative class in Persian like other languages in that the tokens can freely collocate, so they are not considered as MTUs. Types 2, 3 and 4 which are related to the non-compositional meaning, are involved in the MTU discussion. The 2nd type, being the non-compositional construction, builds a closed set and is not generative. This type builds the compound words. These types of MTUs are *static*. That is, the whole phrase is considered as one unit with one morphological or syntactic tag. The internal syntactic structure may be hidden in some compound words like compound conjunctions and adverbs, but they are not hidden in complex predicates (section 4.1). The 3rd type that is related to the compositional and non-compositional constructions is the most problem raising type in tagging. In Peykare, this type is not considered as an MTU, so the tokens are POS-tagged independently. In this type, the frequency of the compositional construction is usually more than the non-compositional one. To bracket the non-compositional construction, we can consider the phrase as the compositional case and in addition to that, assign a semantic tag to the whole phrase. For instance in (2), the adverb [be sæXti] is tagged the same as PP with a (-MNR) semantic tag: [PP-MNR be sæXti]. This type constitutes an open set and it is possible to define some templates for it. So they create the *dynamic MTUs*. In the cases where the non-compositional construction has more frequency, although the MTU is generative and dynamic, it is preferred to assign a single POS tag to the phrase and hide its internal structure.

For tagging the 4th type which is semi-compositional, we have two options. The first option is to consider the phrase as one word and assign it a POS tag, e.g., سه نفره, [se næfære], "threesome, by/for three persons", is regarded as a single word with an ADJ or ADV tag. The second option is regarding the tokens as one phrase with a function tag. The first option is not suitable since it is in opposition with the language generativeness. The second option is preferred since it creates a dynamic and so a generative MTU.

As we can see, not being able to segment the text by correctly phrasing the MTUs causes the Persian machine translation systems to confront troublesome problems in finding the correct target language counterparts.

In the next two sections we describe the static and dynamic MTUs in more detail.

4 Static MTUs

In the previous section it was mentioned that the static MTUs build a closed set in Persian and are non-compositional having separate entries in the lexicon. This class of MTUs (except CPRs) by being non-generative and closed do not pose grave problems for machine translation systems. In this section we describe these MTUs.

4.1 Complex Predicates(CPRs)

The linguists and language engineers confront serious problems in the analysis of Persian CPR in Lexicalist frameworks and need to claim that Persian CPRs are instances of *idioms*, receiving a separate entry in the lexicon complete with their syntactic structure (Karimi, 2005). Persian CPR cannot be considered a lexical unit since its non-verbal (NV) element and light verb (LV) may be separated by a number of elements (Karimi, 1992). In Folli et al. (2005), it is argued that the conflicting properties of Persian CPR can be easily

accommodated in a non-Lexicalist theory such as distributed morphology, where all interpretation occurs post-syntactically and employing Hale and Keyser's (1993, 2002) model, the provided structures of this model translate naturally to Persian CPR. For instance, the syntactic structure of an inchoative CPR like بیدار شدن, [bidɑr ʃodan], (awake becoming), "to wake up", quoted from Karimi (2005) is shown in (3).

(3) [vP [AdjP [DP Kimea] [Adj bidar]] [v ʃod]]

The NV element of a CPR can be PP e.g., **به دنیا** آمدن, [**be donjɑ** ɑmædæn], (to world coming), "to be born", Particle, adjective, noun, predicative noun e.g., **شکست** خوردن, [**ʃekæst** Xordæn], (defeat colliding), "to be defeated" and predicative adjective (Karimi, 1997). The verb [ʃodæn] is used in passive and unaccusative structures. Light verbs can also be used as main verbs. The NV element and LV of a CPR are separate tokens in syntactic level. Both the simple and light verbs are annotated as the head of a vP and other constituents are located inside a predicate phrase (PredP) in a vP.

The compound infinitives such as نگاه کردن, [negɑh kærdæn], (look doing), "to look" are tagged as compound nouns. In bracketing, a compound infinitive is tagged as a noun phrase. This is contrary to the English Penn Treebank where the infinitive is tagged as a VP and it gets a syntactic function tag (nominal) (-NOM) (Bies et al., 1995).

Although most of the complex predicates build an open set in Persian, it is possible to define *static MTU verbal templates* for some less frequent CPRs with LVs such as شستن, [ʃostæn], (to wash) and چیدن, [ʧidæn], (to arrange), although it is not recommended.

4.2 Compound Prepositions and Compound Conjunctions

The compound prepositions (CPREPs) and compound conjunctions (CCONJs) are probably the most frequent MTUs in Persian. In languages like English, some words can be used as a preposition (P) or a conjunction (CONJ), e.g., *after, before*. (Bies et al., 1995). In Persian, a number of CPREPs and CCONJs are somehow alike, the CCONJs having the additional words این که, [ʔinke], "this that" or آن که, [ʔɑnke], "that that", e.g., به خاطر این که, [be Xɑter-e ʔin ke], "because". The simple Ps that can be incorporated with nouns and create the CPREPs are [dær], [be], [bɑ], [ʔæz] and [bær]. They do not take genitive ending /-e/ (POS tag: EZ) and having the [-V, -N] features are considered as "true" Ps (Samiian, 1991).

There is not a consensus about the CPREPs. The CPREPs constituted of a P and an N, are similar to PPs, but there are several morphological and syntactic evidences that show they are different (Abolhassani, 2006). One group of CPREPs includes two Ps e.g, [æz bærɑj-e], "for". The other group has two properties: 1) The Ps have large frequency in language and also concatenated forms in Peykare, and 2) they behave differently from similar PPs. For instance, [be væsile-je], "by", [be ʤoz], "except" and [be ʔellæt-e], "because of". These CPREPs must be tagged as one P. Ps in a language build a closed set. In Peykare we found 38 CPREPs, but based on the above-mentioned properties there are about 105 CPREPs in Persian. Some of them e.g., [be ʔellæt-e], "because of", are incorrectly tagged as MUTs (PREP, N,COM,SING,EZ), while they are just Ps.

There are about 130 CCONJs in Persian e.g., [ægær ʧe], "although" and [æz hengɑm-i ke], "since". Multi-token CONJs in Persian are in fact, the subordinating CONJs. We established two structures, *static* and *dynamic*, for these MTUs in Persian. In static structure, the CCONJ is considered as one unit and is located as the head of a CP like the single CONJ [ke], "that", e.g., [CP [bɑ ʔin voʤud [vP…] and [CP ægær ʧe [vP…]]. In section 5.2, we discuss the dynamic multi-token conjunctions.

4.3 Compound Adverbs

Compound adverbs (CADVs) usually consist of one preposition and one or two nouns. It is important to have in mind that if such a collocation is compositional, it is better not to call it a compound adverb, because we have reduced a generative structure to a single morphological category. In Persian, some CADVs have large frequency e.g., [dær vɑɢeʔ], "really" has 1472. There are about 80 CADVs in Peykare, e.g, [be ʔeXtesɑr], "briefly", [be Xubi], "well" and [dær næhɑjæt], "finally". Since a number of PPs can function as adverbs dynamically, we discuss such cases in 5.3.

4.4 Compound Adjectives

One group of multi-token adjectives consists of two tokens, ADJ + N, e.g, گران قیمت, [gerɑn ɢejmæt], "expensive". Some prefixes such as [mijɑn], "mid-", [zir], "under-", attached to other ADJs build compound adjectives (CADJs) and so is the prefix [ʃebh-e], "semi-, -like", e.g, [ʃebh-e do:læt-i], "semi-governmental". Other CADJs consist of ADJs + Past Participles, e.g., [gærmɑ zæde], (heat hit), "suffering from heatstroke". These MTUs are POS tagged correctly in Peykare. The preposition [be] plus some nouns create CADJs, but in Peykare are POS-tagged separately, e.g., [be ʤɑ], (to place), "suitable".

By the analysis of Pekare, it was found that the MTUs building a closed set of adjectives and being used in non-compositional constructions are taken as CADJ and POS-tagged correctly. Many adjectives are formed generatively by the bound prefixes [bɑ-], "-ful" and [bi-], "-less", e.g., [bɑ deɢɢæt], "careful". These MTUs get the ADJ tag in Peykare. In section 5.4, dynamic adjectival MTUs are discussed.

4.5 Compound Nouns

Another group of MTUs is the compound nouns (CNs). They are POS-tagged correctly in Peykare, most of the times.

The compound nouns in Persian may be neither closed nor static. In section 5.5, we have analyzed some of them.

5 Dynamic MTUs

Most of the Multi-token words being as compound categories are treated as one unit with one POS tag, and their internal syntactic structures are usually hidden. In this section, we discuss the MTUs which constitute an open and generative class in Persian and it is possible to define particular templates for them. By the definition of such templates, the parsers can mark the syntactic phrases more precisely so that the Persian language machine translation systems can use the syntactic and semantic tags to find the exact target language counterparts.

5.1 Verbs and Infinitives

To define such templates, for each LV in the lexicon database, the NV elements are extracted and inserted in tables which are related as one-to-many to the LV tables. For instance, the template for the verb [ʔɑværdæn] which is derived from Peykare is shown in (4).

(4) [vP [PredP x] [v ʔɑværdan]
x ɛ {N, Adj, Particle, PP}
N={ʔɑb, bɑr, dælil, ʤɑ, hoʤum, forud, pædid, færɑham, ræʔj, tɑb, dævɑm, vɑred, feʃɑr, ʔimɑn, ʔozr, bæhɑne, ruj, tæʃrif, ræhm, dævɑm}
Adj={gerd, kæm}
Particle={bɑz, bær, dær, bær, birun, pɑjin, piʃ, foru}
PP={ be ʤɑ(j), be ʧæng, be hesɑb, be Xɑter, be Xæʃm, be Xod, be dær, be dærd, be dæst, be ruj-e kɑr, be zæbɑn, be ʃomar, be ʃur, be sæhne, be ʔæmæl, be kæf, be mijɑn, be næzær, be væʤd, be voʤud, be hæmrɑh, be huʃ, be jɑd}

These templates can help the Persian machine translation systems in the segmentation of the complex predicates.

As we saw in section 4.1, the multi-token verbs and infinitives are the CPRs and most of them constitute an open set in Persian, so they are somehow generative. An LV such as کردن, [kærdæn], has completely lost its main verb usage and is used vastly in making verbs of the new imported nouns like *fax, email, post, format* and *click*, e.g., فکس کردن, [fæks kærdan], "to fax".

The other dynamic MTU verbal template that can be defined is the passive construction with the auxiliary verb شدن, [ʃodæn], "to become". These constructions can also be considered as CPR (Karimi, 2005). Therefore, the past participle, derived from a transitive verb, together with [ʃodæn], builds a dynamic MTU vebal template, e.g., دیده شده اند, [dide ʃode ʔænd], "they have been seen".

5.2 Prepositions and Conjunctions

The tokens such as [(ɢæbl/piʃ) æz], "before" and [(bæʔd/pæs) æz)], "after", cannot be considered as one unit, because their internal constituents can be coordinated, e.g., [ɢæbl va bæʔd az zohr], "before and after noon". Contrary to the Penn English Treebank where all the multi-word prepositions such as *because of*, have flat structures and are considered as a single P, we regard this kind of multi-token prepositions as two PPs (5).

(5) قبل از انقلاب
[PP-TMP ɢæbl [PP ʔæz [NP ʔenɢelɑb]]]

"before the revolution"

Since the multi-token prepositions are static and build a closed set, we cannot define a useful *MTU prepositional template*. For the *MTU conjunctional template* about 40 compound prepositions can collocate with [ʔin ke], "this that" and [ʔɑn ke], "that that" and create conjunctions. Also, some PPs plus the conjunction [ke], "that" build conjunctions. In bracketing, they are considered as complements of the PPs (6).

(6) به این علت که or به علت اینکه
[PP-PRP be ʔin ʔellæt [CP ke [vP …]]]
[PP-PRP be ʔellat-e ʔin [CP ke [vP …]]]
"because"

These subordinating conjunctions can be used in sentential/verbal adjuncts, adjuncts or complements of nouns, predicates, complements of PPs and CPREPs.

5.3 Adverbs

In this section, we define and analyze the *MTU adverbial templates*. Many PPs building the dynamic MTUs can function as adverbs. One template consists of PPs having a PREP and a N,COM,SING, e.g., [be nærmi], "gently", [bɑ ʃur o ʃo:ɢ], "enthusiastically". These are adverbials of manner, so they are bracketed as PPs with a semantic function tag (-MNR) e.g., [PP-MNR be nærmi], but if they are written as concatenated, they are tagged as an adverbial phrase, e.g., بەنرمی,[ADVP-MNR benærmi]. As mentioned before, these constructions can also be compositional. The POS tags in Peykare do not help us to determine the semantic tags precisely. So, a manual correction is needed in Persian Treebank generation.

Another template consists of PPs with manner function. The tokens include [be (to:r-e/tærz-e/ʃekl-e/gune-je/no:ʔ-e/suræt-e/ʃive-je)], "in a … way/manner", and an adjectival phrase, e.g., [be ʃekl-e Xɑreɢolɑde-ʔi], "extraordinarily". These are also bracketed as PPs with a semantic function tag (-MNR).

5.4 Adjectives

One of the MTU adjectival templates is related to the composition of prepositions as prefixes and other constituents. One group consists of prefixes such as [æz piʃ/ɢæbl], "pre-" plus N plus the past participle [ʃode], "become", e.g., از قبل تعیین شده, [æz ɢæbl tæʔjin ʃode], "predetermined". Another prefix is [æz Xod], "self", that together with another adjective builds a compound adjective, e.g., از خود راضی, [æz Xod rɑzi], "self-satisfied". When these tokens are written as concatenated, they receive ADJ tag in Peykare. In bracketing, we analyzed these MTUs as simple adjectives.

As mentioned in 3.2, one of the most generative compounds is related to the semi-compositional constructions which consist of a number (NUM) plus a residual (RES).

Another group consists of prefixes [ɢɑbel-e], "-able, -ible" and [ɢejr-e ɢɑbel-e], "un-, -in___-able, -ible", e.g., [ɢejr-e ɢɑbel-e bɑvær], "unbelievable".

Some adjectives are formed by adding the adjective-maker suffix [-i] to the infinitive, e.g., از بین رفتنی, [æz bejn ræftæn-i], "destructible".

5.5 Nouns

As mentioned in 4.5, most of the multi-token nouns were tagged correctly in Peykare. In this section we define some of the *MTU nominal templates*. Some compound nouns can be very generative. If the second noun is [Xɑne], "house" or [foruʃ], "sale", e.g., دارو خانه,[dɑru Xɑne], "drugstore", کتاب فروش, [ketɑb foruʃ], "bookseller". Also CNs consisting of a preposition as prefix plus a noun or the past/present stem, e.g., پیش پرداخت, [piʃ pærdɑXt], "prepayment", پس لرزه, [pæs lærze], "aftershock".

Some are less generative such as when the first noun is always without genitive ending [-e] (kæsre-ezɑfe)(POS tag: EZ) and the second noun may function as a subjective suffix, e.g., [dɑr], "-ist, owner", e.g., داروخانه دار, [dɑruXɑne dɑr], "pharmacist".

Some compound nouns are generated from compound adjectives plus noun-maker suffix [-i], e.g., از خود گذشتگی, [æz Xod gozæʃte-gi], "selflessness". These MTUs are considered as one morphological category.

So there is no need to add these CNs in the lexicon if we have a good analyzer and these compounds are compositional.

6 Conclusion

In this paper, we analyzed the morphological and syntactic annotation of the MTUs. Due to the

concatenation nature of the Persian script, the orthographic variation (Lazard, 1992; Buckwalter, 2004) and the existence of both compositional and non-compositional constructions for some tokens that results in different POS and syntactic tags, and also the semi-compositional constructions for tokens, two kind of MTUs, static and dynamic, were defined and applied to the verbs, infinitives, prepositions, conjunctions, adverbs, adjectives and nouns.

The static MTUs are related to the non-generative and non-compositional constructions that comprise closed sets. They are normally tagged as one category, except for CPRs (Karimi, 2005; Folli et al., 2005). In bracketing the generative compositional constructions, in order not to reduce the syntactic phrases to morphological categories, the MTU templates are defined wherever possible, i.e., the compositional construction of a phrase is kept unchanged most of the times, and the whole phrase receives a syntactic or semantic function tag. However, in some compositional and semi-compositional constructions, the whole phrase gets one POS tag, so inevitably its internal structure is hidden.

The machine translation systems using the rule-based methods to parse the texts can also utilize the results which are directly related to text segmentation. Also we saw that how the approaches to dealing with MTUs, influence the generation of Persian Treebanks (Bies et al., 1995; Bies and Maamouri, 2003; Marcus et al., 1993; Santorini and Marcinkiewicz, 1991; Sima'an et al, 2001).

References

Ann Bies, Mark Ferguson, Karen Katz, and Robert MacIntyre. 1995. *Bracketing Guidelines for Treebank II Style: Penn Treebank Project* ftp://ftp.cis.upenn.edu/pub/treebank/doc/manual

Ann Bies and Mohamed Maamouri. 2003. *Penn Arabic Treebank Guidelines.* Linguistic Data Consortium. University of Pennsylvania.

Beatrice Santorini and Mary Ann Marcinkiewicz. 1991. *Bracketing guidelines for the Penn Treebank Project.* Ms., Department of Computer and Information Science, University of Pennsylvania. ftp://ftp.cis.upenn.edu/pub/treebank/doc/cl93.ps.gz

Geoffry Leech, and A. Wilson. 1996. *EAGLES Recommendations for the Morphosyntactic Annotation of Corpora.* EAGLES-Guidelines EAG--TCWG--MAC/R. Final version of 3.1996. EAGLES, Instituto di Linguistica Computazionale, Pisa.

Gilbert Lazard.1992. *A Grammar of Contemporary Persian*. Mazda Publishers, Costa Mesa, California.

Jan Cloeren. 1999. *In Syntactic Wordclass Tagging*. In Hans van Halteren. Dordrecht: Kluwer Academic Publishers.

Jan Mohammad and Simin Karimi. 1992. *Light verbs are taking over: Complex verbs in Persian*. Proceedings of WECOL, pp. 195-212.

Karine Megerdoomian. 2000. *Unification-Based Persian Morphology*. In Proceedings of CICLing.

Kenneth Hale and Samuel Keyser. 1993. *On argument structure and the lexical expression of syntactic relations*. In: Hale, K., Keyser S.J., (Eds.), View from Building 20. MIT Press, Cambridge, MA, pp. 53-109.

Kenneth Hale and Samuel Keyser. 2002. *Prolegomenon to a Theory of Argument Structure.* Cambridge/London: The MIT Press.

Mahmood Bijankhan, Javad Sheykhzadegan and Mohammad Bahrani. 2008. *Lessons From Designing A Persian Resource: Peykare.* (Under review)

Mitchel P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz . 1993. *Building a large annotated corpus of English: The Penn Treebank*. Computational Linguistics 19(2), 313-330, 1993.

Noam Chomsky and Morris Halle. 1968. *The sound pattern of English*. New York: Harper and Row, Boston: MIT Press.

Paul L. Garvin. 1954. *Delimitation of Syntactic Units*. in Language, vol. 30, no. 3, pp. 345-348, Linguistic Society of America.

Persian Academy of Language and Literature. 2005. *Persian Orthography.* http://www.persianacademy -.ir/fa/dastoorpdf.aspx

Rafaella Folli, Heidi Harley, and Simin Karimi. 2005. *Determinants of event type in Persian complex predicates*. Lingua.

Simin Karimi. 1997. *Persian Complex Verbs: Idiomatic or Compositional*. Lexicology 3, 273-318.

Simin Karimi. 2005. *A minimalist approach to Scrambling*. Mouton de Gruyter, Berlin/New York.

Tim Buckwalter. 2004. *Issues in Arabic Orthography and Morphology Analysis*. In Proceedings of COLING.

Vida Samiian. 1991. *Prepositions in Persian and the Neutralization Hypothesis*. California State University, Fresno.

Willem L. Graff. 1929. *The Word and the Sentence*. in Language, vol. 5, no. 3, pp. 163-188, Linguistic Society of America.

Zahra Abolhassani. 2006. *An Account for Compound Prepositions*. Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 113–119, Sydney.