Research on the Automatic Extraction of Persian Simple Sentences

Wei Li

Luoyang University of Foreign Languages471003 Luoyang, Henan, China

lw_waiyuan@163.com

Keywords: Persian; Corpus; Clause; Automatic extraction

Abstract. In order to meet the needs of natural language shallow parsing for Persian, this paper analyzes the boundary rules of Persian sentences and puts forward an effective algorithm and program of extracting Persian simple sentences by repeated experiments on the basis of Uppsala Persian Corpus. This algorithm, which can reduce the difficulty of parsing, is of vital importance to information retrieval, machine translation, phrase recognition, the construction of network answering system, etc. Besides, the obtained simple sentences can also be useful resources for Persian linguistic studies.

Introduction

One of the natural language shallow parsing tasks is resolving compound sentences into simple ones, which can lower the difficulty of parsing significantly through simplifying sentence structures. Nowadays, some achievements have been made in the aspect of Persian part-of-speech tagging and there are not only part-of-speech tagging systems of high accuracy, but well-known Persian corpora like Bijankhan Corpus [1] and Uppsala Persian Corpus (hereinafter "UPC") [2] have been constructed as well. According to the corpora, there are a lot of complex sentences in Persian texts. If the complex sentences can be divided into simple sentences, it is of great significance to both improving processing efficiency and deepening the research on Persian corpora. Therefore, this paper explores characteristics of Persian complex sentences and designs a system of extracting Persian simple sentences on the basis of UPC.

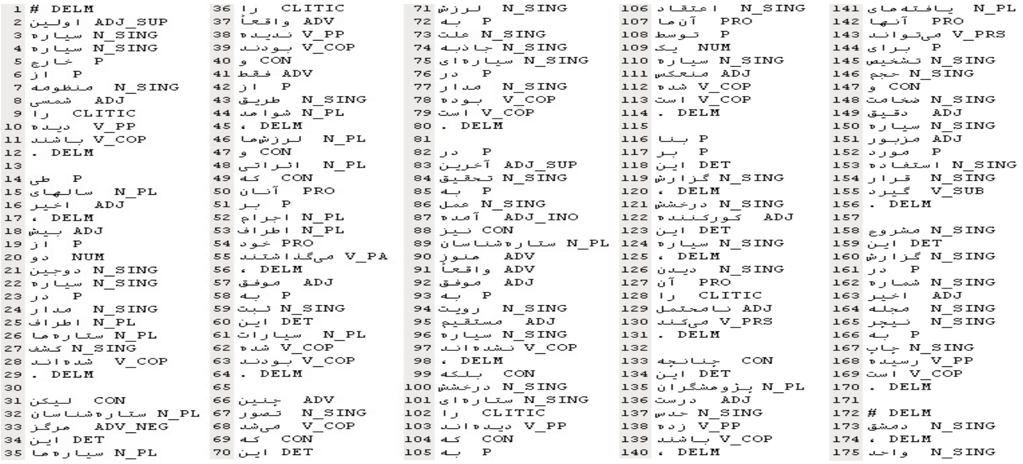
Related Work

Because of the significance of recognizing simple sentences automatically, a lot of researchers at home and abroad have undertaken studies on this aspect. For example, some researchers have discussed the end forms of the Uyghur sentences and realized automatic identification of Uyghur simple sentences by means of the corpus [3]. Some other researchers consider that the clause is the basic unit of Chinese discourses and have implemented a comma-based system of identifying clauses via tagged corpus [4]. Another two scholars have undertaken some studies on topic detection of punctuation sentences by means of the generalized topic theory in order to improve the accuracy of information extraction and machine translation [5]. Still another three think that processing complex sentences correctly in English-Chinese machine translation depends on whether the clauses can be identified and have put forward the detection method of English clauses based on maximum entropy principle [6]. Also, another detection method of English clauses has been discussed in [7] and improvement proposals of adding linguistic rules to the model and modifying smoothing algorithm have been made in the conclusion. [8] has discussed a memory-based method of extracting English clauses, which needs the part-of-speech tags of the words and a base chunk structure of the sentence.

To sum up, simple sentence identification is of great importance in natural language processing and the study of linguistics.

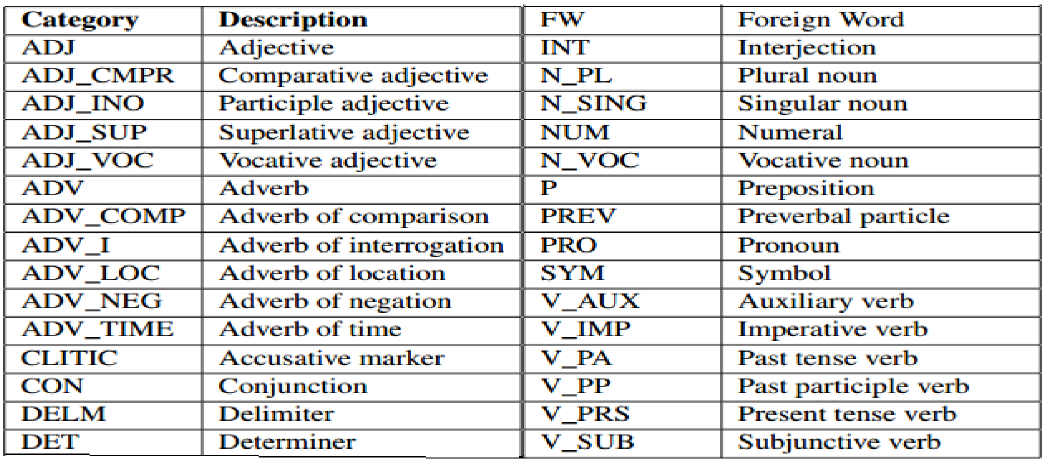
Uppsala Persian Corpus

UPC is the improved version of Bijankhan Corpus. Bijankhan Corpus was released in 2004, which is the first large-scale manually annotated corpus. Its contents are from online novels and newspapers, etc and have the features of various themes and rich topics. All the words in Bijankhan Corpus are tagged, which contains about 2.6 million tokens. However, as it was constructed not for natural language processing, many features make it not suitable for automatic processing. For example, there exist the problems of lacking sentence segmentation and uneven normalization. Besides, Persian and Arabic letters are mixed together and texts have a mixture of different types of character sets. Therefore, Mojgan Seraji improved Bijankhan Corpus and named the new version UPC. There are two improvements in UPC. For one thing, the tagset of Bijankhan Corpus has been modified via removing, adding and merging categories. For another, mistakes and inconsistencies in the annotation have been corrected. UPC is available at [9]. A screenshot of UPC and its tagset are shown in Fig. 1 and Fig. 2,



```
# DELM
اولین ADJ_SUP
سیاره N_SING
سیاره N_SING
خارج P
از P
منظومه N_SING
شمسی ADJ
را CLITIC
دیده V_PP
باشند V_COP
. DELM

طی P
سالهای N_PL
اخیر ADJ
، DELM
بیش ADJ
از P
دو NUM
دوجین N_SING
سیاره N_SING
در P
مدار N_SING
اطراف N_PL
ستارهها N_PL
کشف N_SING
شدهاند V_COP
. DELM

لیکن CON
ستارهشناسان N_PL
هرگز ADV_NEG
این DET
سیارهها N_PL
را CLITIC
واقعاً ADV
ندیده V_PP
بودند V_COP
و CON
فقط ADV
از P
طریق N_SING
شواهد N_PL
، DELM
لرزشها N_PL
و CON
اثراتی N_PL
که CON
آنان PRO
بر P
اجرام N_PL
اطراف N_PL
خود PRO
میگذاشتند V_PA
، DELM
موفق ADJ
به P
ثبت N_SING
این DET
سیارات N_PL
شده V_COP
بودند V_COP
. DELM

چنین ADV
تصور N_SING
میشد V_COP
که CON
این DET
لرزش N_SING
به P
علت N_SING
جاذبه N_SING
ستارهای N_SING
در P
مدار N_SING
بوده V_COP
است V_COP
. DELM

در P
آخرین ADJ_SUP
تحقیق N_SING
به P
عمل N_SING
آمده ADJ_INO
نیز CON
ستارهشناسان N_PL
هنوز ADV
واقعاً ADV
موفق ADJ
به P
رویت N_SING
مستقیم ADJ
سیاره N_SING
نشدهاند V_COP
، DELM
بلکه CON
درخشش N_SING
ستارهای N_SING
را CLITIC
دیدهاند V_PP
که CON
به P
اعتقاد N_SING
آنها PRO
توسط P
یک NUM
سیاره N_SING
منعکس ADJ
شده V_COP
است V_COP
. DELM

بنا P
بر P
این DET
گزارش N_SING
، DELM
درخشش N_SING
کورکننده ADJ
این DET
سیاره N_SING
، DELM
دیدن N_SING
آن PRO
را CLITIC
نامحتمل ADJ
میکنند V_PRS
. DELM

چنانچه CON
این DET
پژوهشگران N_PL
درست ADJ
حدس N_SING
زده V_PP
باشند V_COP
، DELM
یافتههای N_PL
آنها PRO
میتوانند V_PRS
برای P
تشخیص N_SING
حجم N_SING
و CON
ضخامت N_SING
دقیق ADJ
سیاره N_SING
مزبور ADJ
مورد P
استفاده N_SING
قرار N_SING
گیرد V_SUB
. DELM

مشروع N_SING
این DET
گزارش N_SING
در P
شماره N_SING
اخیر ADJ
مجله N_SING
نیچر N_SING
به P
چاپ N_SING
رسیده V_PP
است V_COP
. DELM

# DELM
دمشق N_SING
، DELM
واحد N_SING
```

Figure 1. A screenshot of UPC



Category	Description	FW	Foreign Word
ADJ	Adjective	INT	Interjection
ADJ_CMPR	Comparative adjective	N_PL	Plural noun
ADJ_INO	Participle adjective	N_SING	Singular noun
ADJ_SUP	Superlative adjective	NUM	Numeral
ADJ_VOC	Vocative adjective	N_VOC	Vocative noun
ADV	Adverb	P	Preposition
ADV_COMP	Adverb of comparison	PREV	Preverbal particle
ADV_I	Adverb of interrogation	PRO	Pronoun
ADV_LOC	Adverb of location	SYM	Symbol
ADV_NEG	Adverb of negation	V_AUX	Auxiliary verb
ADV_TIME	Adverb of time	V_IMP	Imperative verb
CLITIC	Accusative marker	V_PA	Past tense verb
CON	Conjunction	V_PP	Past participle verb
DELM	Delimiter	V_PRS	Present tense verb
DET	Determiner	V_SUB	Subjunctive verb

Figure 2. The tagset of UPC

An Algorithm of Extracting Persian Simple Sentences Based on Grammar Rules

According to the Persian grammar, sentences with two or more predicate verbs are called complex sentences, which occupy an important position in Persian expressions. According to the different relationship between clauses, complex sentences can be divided into compound complex sentences and main-subordinate complex sentences. This paper develops the following steps of extracting simple sentences based on the idea of filtering:

First, punctuation marks are applied to make a rough split of complex sentences. In UPC, all the punctuation marks are annotated with “<DELM>”. It can be found by traversing the corpus that clauses are joined mainly by comma, colon and semicolon, which appear as “،”, “:” and “؛” in Persian.

Through traversing the corpus, I extract every clause joined by comma together with a verb, colon and semicolon and save them as a single line respectively.

Second, lines with no verb are deleted and lines with only one verb are saved as the target simple sentences. The rest are further processed.

Persian modal verbs like “بایستن(should)”, “شایستن(may)” have gradually been turned into adverbs, which are represented in forms of verbs’ not changing with the subject [10]. The usual forms of “بایستن(should)” and “شایستن(may)” are “شاید”, “باید”, “بایستی”, “می بایست”, “بایست”. Furthermore, when “توانستن(can)” and “شدن(can)” are used to make an impersonal sentence, their forms also don’t vary with the subject and they have only three forms “بتوان”, “میتوان” and “میشود”. In the corpus, all the words mentioned above are annotated as “<V_AUX>”, which means auxiliary verb. Actually, they function as adverbs. Therefore, verbs with “<V_AUX>” are not regarded as a verb. Moreover, two verbs are often joined together in Persian and the form in the corpus is “word<V_? > word<V_? >”. As they function as predicate verbs together, two verbs which are next to each other are regarded as one verb. As for “و<CON>(and)”, “یا<CON>(or)” and other conjunctions, if they join two or more verbs, all of them are regarded as one verb.

Third, in this step, a decision should be made which should be processed first, compound complex sentences or main-subordinate complex sentences. The following three examples are used to make an illustration.

(1)... و پرویز هم پسری کوشا ست که هر روز صبح به میدان ورزش می رود و همکلاسهای دیگر باید از آنها تحصیل می کنند....

···and Parviz is also a diligent boy, who goes to the playground early in the morning every day and the other classmates should learn from him…

(2)چگونه می توان به کودکی که نه می شنود و نه حرف می زند خواندن و نوشتن آموخت؟

How can we teach the children who are deaf and dumb to read and write?

(3)همه کشورها اعم از اینکه بزرگ و کوچک باشند، باید متساوی الحقوق باشند.

All countries, big or small, should be equal.

In Example (1), there is a subordinate clause in a compound complex sentence. “و” is a coordinate conjunction, which is the equal of “and” in English. “که” is the relative pronoun of the attributive clause, which is the equal of “that”. If the coordinate conjunction “و” is processed first and I extract compositions until “است”, the clause led by “که” will be separated and the pattern “noun+verb+clause” cannot be found. In Example (2), there is a coordinate clause in a main-subordinate complex sentence. If the clause led by “که” is extracted first, the verb pattern won’t be broken in spite of separating the coordinate clauses from each other. Also, the usage of the conjunction “و” is very common in Example (3). If “و” is processed first, the original structure of the sentence will be broken. Therefore, main-subordinate complex sentences in the corpus should be processed first.

In Persian, main-subordinate complex sentences fall roughly into three categories: noun clauses, adjectival clauses and adverbial clauses. Noun clauses contain subject clauses, object clauses, appositive clauses and predicative clauses. Subject clauses can be divided into preposition subject clauses and post subject clauses according to different positions of the clause, as are shown in Example (4) and Example (5). As Persian belongs to the type of SOV, the verb is preceded by the object. However, there are two situations about the position of object clauses, which can be located either before the verb or after the verb, as are shown in Example (6) and Example (7). Predicative clauses usually come after link verbs, as is shown in Example (8). Adjectival clauses refer to attributive clauses, which are similar to appositive clauses. The difference is that the relative pronoun of attributive clauses is used as a sentence element and the relative pronoun of appositive clauses not. The position of adverbial clauses is very flexible, which can be placed in the initial, medial and end position.

(4)هر چه می درخشد طلا نیست.

All is not gold that glitters.

(5)معلوم نیست که ساعت چند بر می گردد.

It is not clear when he will be back.

(6)ما باید آنچه را که تحصیل می کنیم باهم صحبت کنیم.

We should discuss together what we learn.

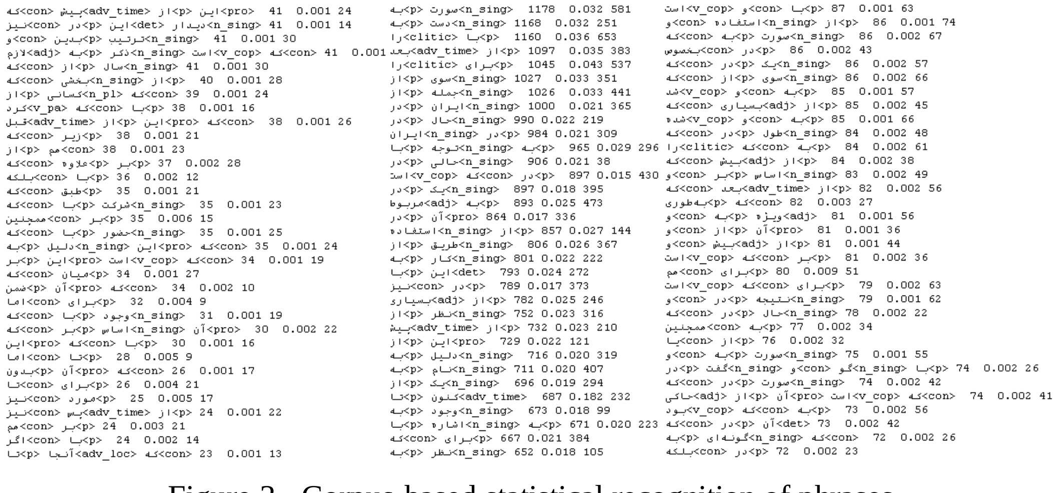
(7)می دانید که این بازی چه موقع شروع می شود؟

Do you know when this game will start?

(8)نخستین کار برنامه امشبم این است که این کلمات را حفظ می کنم.

My first study plan this evening is remembering all the words.

In Example (4), "هر چه می درخشد(that glitters)" is the subject clause, which is in the initial position and belongs to preposition subject clauses. The relative pronouns are usually "هر چه", "آنچه", "آنچه که", "هر چه که", "آنکه", "هر کس", "هر که", "اینکه". Example (6) belongs to the situation that the object clause comes before the verb and its relative pronouns share the same ones above. Example (5) and Example (7) stand for post subject clauses and post object clauses, which share the same complementizer "که" with appositive clauses, predicative clauses and attributive clauses. "where", "when", etc in English can be used as relative adverbs and "چه موقع" in Example (7) is the equal of "when" in English, which is not the relative adverb and its position is usually medial. In order to keep as many original sentence elements as possible, take the clause led by "که" for example, sentence elements from "که" to the first verb behind "که" are extracted as a simple sentence and if there is still a verb in the rest of the original sentence, "<C>" is added in the missing position, which represents that a clause can be added here. This method can be named Boundary Word Interception. If the relative pronoun "که" is omitted, the sentence should be disconnected from behind the verb. This method can be named Verbal Disconnection. Here, words like "که" are named the boundary words, which need to contain the part-of speech tagging information, because sometimes a word has different parts of speech. For instance, "تا<CON>" is a complementizer, while "تا<P>" is a preposition. "تا<CON>" should be used to extract simple sentences instead of "تا<P>". Moreover, there are two tagging forms about the multi-word boundary word in the corpus. For example, "قبل‌ازاینکه<CON>" and "پیش<ADV_TIME> از<P> آن<PRO> که<CON>" both appear in the corpus. In order to identify all the boundary words, a corpus-based statistical recognition method is first applied and the result of recognition is shown in Fig. 3 and the three parameters behind the phrases are phrase frequency, phrase probability and the minimum value of number of different words in either the left or the right neighboring place of the possible phrase.



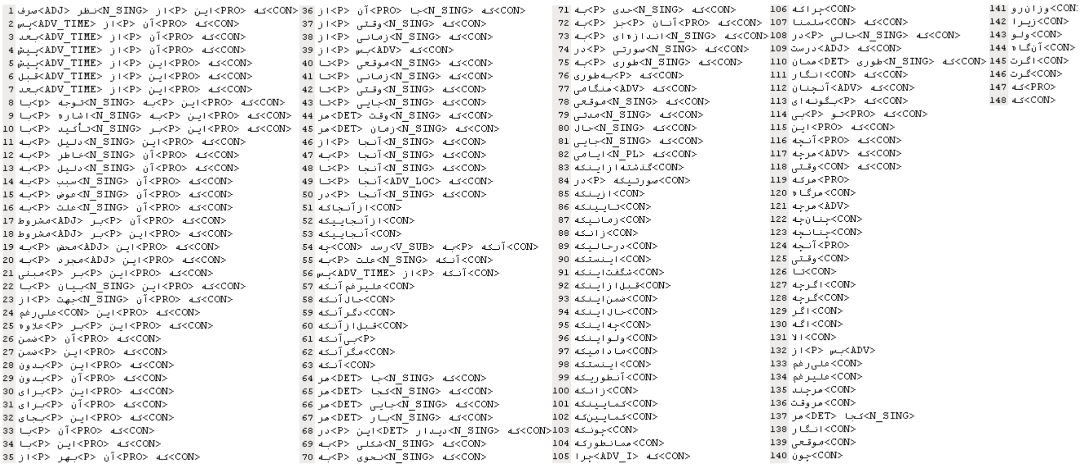
```
که<con> پیش<adv_time> از<p> این<pro>  41  0.001 24
نیز<con> در<p> این<det> دیدار<n_sing> 41  0.001 14
و<con> بدین<p> ترتیب<n_sing>  41  0.001 30
لازم<adj> به<p> ذکر<n_sing> است<v_cop> که<con> 41  0.001
که<con> از<p> سال<n_sing> 41  0.001 30
که<con> بخشی<n_sing> از<p>  40  0.001 28
از<p> کسانی<n_pl> که<con> 39  0.001 24
کرد<v_pa> که<con> با<p> 38  0.001 16
قبل<adv_time> از<p> این<pro> که<con>  38  0.001 26
که<con> زیر<p>  38  0.001 21
از<p> هم<con> 38  0.001 23
که<con> علاوه<p> بر<p> 37  0.002 28
بلکه<con> با<p> 36  0.002 12
که<con> طبق<p>  35  0.001 21
که<con> با<p> شرکت<n_sing>  35  0.001 23
همچنین<con> بر<p> 35  0.006 15
که<con> با<p> حضور<n_sing>  35  0.001 25
به<p> دلیل<n_sing> این<pro> که<con> 35  0.001 24
بر<p> این<pro> است<v_cop> که<con> 34  0.001 19
که<con> میان<p> 34  0.001 27
ضمن<p> آن<pro> که<con>  34  0.002 10
اما<con> برای<p>  32  0.004 9
که<con> با<p> وجود<n_sing>  31  0.001 19
که<con> بر<p> اساس<n_sing> آن<pro>  30  0.002 22
این<pro> که<con> با<p>  30  0.001 16
اما<con> تا<p>  28  0.005 9
بدون<p> آن<pro> که<con> 26  0.001 17
تا<con> برای<p> 26  0.004 21
نیز<con> مورد<p>  25  0.005 17
نیز<con> پس<adv_time> از<p> 24  0.001 22
هم<con> بر<p> 24  0.003 21
اگر<con> با<p>  24  0.002 14
تا<p> آنجا<adv_loc> که<con> 23  0.001 13
```

```
به<p> صورت<n_sing>  1178  0.032 581
به<p> دست<n_sing> 1168  0.032 251
را<clitic> با<p>  1160  0.036 653
بعد<adv_time> از<p> 1097  0.035 383
را<clitic> برای<p>  1045  0.043 537
از<p> سوی<n_sing> 1027  0.033 351
از<p> جمله<n_sing>  1026  0.033 441
در<p> ایران<n_sing> 1000  0.021 365
در<p> حال<n_sing> 990 0.022 219
ایران<n_sing> در<p> 984 0.021 309
با<p> توجه<n_sing> به<p>  965 0.029 296
در<p> حالی<n_sing>  906 0.021 38
است<v_cop> که<con> در<p>  897 0.015 430
در<p> یک<n_sing>  897 0.018 395
مربوط<adj> به<p>  893 0.025 473
در<p> آن<pro> 864 0.017 336
استفاده<n_sing> از<p> 857 0.027 144
از<p> طریق<n_sing>  806 0.026 367
به<p> کار<n_sing> 801 0.022 222
با<p> این<det>  793 0.024 272
نیز<con> در<p>  789 0.017 373
بسیاری<adj> از<p> 782 0.025 246
از<p> نظر<n_sing> 752 0.023 316
پیش<adv_time> از<p> 732 0.023 210
از<p> این<pro>  729 0.022 121
به<p> دلیل<n_sing>  716 0.020 319
به<p> نام<n_sing> 711 0.020 407
از<p> یک<n_sing>  696 0.019 294
تا<p> کنون<adv_time>  687 0.182 232
به<p> وجود<n_sing>  673 0.018 99
با<p> اشاره<n_sing> به<p> 671 0.020 223
که<con> برای<p> 667 0.021 384
به<p> نظر<n_sing> 652 0.018 105
```

```
است<v_cop> و<con> با<p> 87  0.001 63
و<con> استفاده<n_sing> از<p>  86  0.001 74
که<con> به<p> صورت<n_sing>  86  0.002 67
بخصوص<con> در<p>  86  0.002 43
که<con> در<p> یک<n_sing>  86  0.002 57
که<con> از<p> سوی<n_sing> 86  0.002 66
شد<v_cop> و<con> به<p>  85  0.001 57
که<con> بسیاری<adj> از<p> 85  0.002 45
شده<v_cop> و<con> به<p> 85  0.001 66
که<con> در<p> طول<n_sing> 84  0.002 48
را<clitic> که<con> به<p>  84  0.002 61
که<con> پیش<adj> از<p>  84  0.002 38
و<con> بر<p> اساس<n_sing> 83  0.002 49
که<con> بعد<adv_time> از<p> 82  0.002 56
به‌طوری<p> که<con> 82  0.003 27
و<con> به<p> ویژه<adj>  81  0.001 56
و<con> از<p> آن<pro>  81  0.001 36
و<con> پیش<adj> از<p> 81  0.001 44
است<v_cop> که<con> بر<p>  81  0.002 36
هم<con> برای<p> 80  0.009 51
است<v_cop> که<con> برای<p>  79  0.002 63
و<con> در<p> نتیجه<n_sing>  79  0.001 62
که<con> در<p> حال<n_sing> 78  0.002 22
همچنین<con> به<p> 77  0.002 34
یا<con> از<p> 76  0.002 32
و<con> به<p> صورت<n_sing> 75  0.001 55
در<p> گفت<n_sing> و<con> گو<n_sing> با<p> 74  0.002 26
که<con> در<p> صورت<n_sing>  74  0.002 42
حاکی<adj> از<p> آن<pro> است<v_cop> که<con>  74  0.002 41
بود<v_cop> که<con> به<p>  73  0.002 56
که<con> در<p> آن<det> 73  0.002 42
به<p> گونه‌ای<n_sing> که<con>  72  0.002 26
بلکه<con> در<p> 72  0.002 23
```

Figure 3. Corpus-based statistical recognition of phrases

The statistically based approach above overcomes the subjective shortcomings and all the boundary words in the corpus are listed after manual verification. Because there are many phrases in the boundary words, whole word matching can't be applied in the process of identification. However, some boundary words are part of other boundary words. For example, "ه ک<CON>"can serve as a boundary word alone, but it is also a part of many other boundary words. Therefore, in order to find all the boundary words, the boundary words need to be sorted first and the sort result is shown in Fig. 4,



```
1 صرف<ADJ> نظر<N_SING> از<P> این<PRO> که<CON>
2 پس<ADV_TIME> از<P> آن<PRO> که<CON>
3 بعد<ADV_TIME> از<P> آن<PRO> که<CON>
4 پیش<ADV_TIME> از<P> آن<PRO> که<CON>
5 پیش<ADV_TIME> از<P> این<PRO> که<CON>
6 قبل<ADV_TIME> از<P> این<PRO> که<CON>
7 بعد<ADV_TIME> از<P> این<PRO> که<CON>
8 با<p> توجه<N_SING> به<P> این<PRO> که<CON>
9 با<P> اشاره<N_SING> به<P> این<PRO> که<CON>
10 با<P> تأکید<N_SING> بر<P> این<PRO> که<CON>
11 به<P> دلیل<N_SING> این<PRO> که<CON>
12 به<P> خاطر<N_SING> آن<PRO> که<CON>
13 به<P> دلیل<N_SING> آن<PRO> که<CON>
14 به<P> سبب<N_SING> آن<PRO> که<CON>
15 به<P> عوض<N_SING> آن<PRO> که<CON>
16 به<P> علت<N_SING> آن<PRO> که<CON>
17 مشروط<ADJ> بر<P> آن<PRO> که<CON>
18 مشروط<ADJ> بر<P> این<PRO> که<CON>
19 به<P> محض<ADJ> این<PRO> که<CON>
20 به<P> مجرد<ADJ> این<PRO> که<CON>
21 مبنی<P> بر<P> این<PRO> که<CON>
22 با<P> بیان<N_SING> این<PRO> که<CON>
23 از<P> جهت<N_SING> آن<PRO> که<CON>
24 علی‌رغم<CON> این<PRO> که<CON>
25 علاوه<P> بر<P> این<PRO> که<CON>
26 ضمن<P> آن<PRO> که<CON>
27 ضمن<P> این<PRO> که<CON>
28 بدون<P> این<PRO> که<CON>
29 بدون<P> آن<PRO> که<CON>
30 برای<P> این<PRO> که<CON>
31 برای<P> آن<PRO> که<CON>
32 بجای<P> این<PRO> که<CON>
33 با<P> آن<PRO> که<CON>
34 با<P> این<PRO> که<CON>
35 از<P> بهر<P> آن<PRO> که<CON>
36 از<P> آن<PRO> جا<N_SING> که<CON>
37 از<P> وقتی<N_SING> که<CON>
38 از<P> زمانی<N_SING> که<CON>
39 از<P> بس<ADV> که<CON>
40 تا<P> موقعی<N_SING> که<CON>
41 تا<P> زمانی<N_SING> که<CON>
42 تا<P> وقتی<N_SING> که<CON>
43 تا<P> جایی<N_SING> که<CON>
44 هر<DET> وقت<N_SING> که<CON>
45 هر<DET> زمان<N_SING> که<CON>
46 از<P> آنجا<N_SING> که<CON>
47 به<P> آنجا<N_SING> که<CON>
48 تا<P> آنجا<N_SING> که<CON>
49 تا<P> آنجا<ADV_LOC> که<CON>
50 در<P> آنجا<N_SING> که<CON>
51 ازآنجاکه<CON>
52 ازآنجاییکه<CON>
53 آنجاییکه<CON>
54 چه<CON> رسد<V_SUB> به<P> آنکه<CON>
55 به<P> علت<N_SING> آنکه<CON>
56 پس<ADV_TIME> از<P> آنکه<CON>
57 علیرغم آنکه<CON>
58 حال آنکه<CON>
59 دگرآنکه<CON>
60 قبل ازآنکه<CON>
61 بی‌آنکه<P>
62 مگرآنکه<CON>
63 آنکه<CON>
64 هر<DET> جا<N_SING> که<CON>
65 هر<DET> کجا<N_SING> که<CON>
66 هر<DET> جایی<N_SING> که<CON>
67 هر<DET> بار<N_SING> که<CON>
68 در<P> این<DET> دیدار<N_SING> که<CON>
69 به<P> شکلی<N_SING> که<CON>
70 به<P> نحوی<N_SING> که<CON>
71 به<P> حدی<N_SING> که<CON>
72 به<P> جز<P> آنان<PRO> که<CON>
73 به<P> اندازه‌ای<N_SING> که<CON>
74 در<P> صورتی<N_SING> که<CON>
75 به<P> طوری<N_SING> که<CON>
76 به‌طوری<P> که<CON>
77 هنگامی<ADV> که<CON>
78 موقعی<N_SING> که<CON>
79 مدتی<N_SING> که<CON>
80 حال<N_SING> که<CON>
81 جایی<N_SING> که<CON>
82 ایامی<N_PL> که<CON>
83 گذشته ازاینکه<CON>
84 در<P> صورتیکه<CON>
85 ازینکه<CON>
86 تااینکه<CON>
87 زمانیکه<CON>
88 زانکه<CON>
89 درحالیکه<CON>
90 اینستکه<CON>
91 شگفت اینکه<CON>
92 قبل ازاینکه<CON>
93 ضمن اینکه<CON>
94 حال اینکه<CON>
95 چه اینکه<CON>
96 ولواینکه<CON>
97 مادامیکه<CON>
98 اینستکه<CON>
99 آنطوریکه<CON>
100 زانکه<CON>
101 کمااینکه<CON>
102 کمااین‌که<CON>
103 چونکه<CON>
104 همانطورکه<CON>
105 چرا<ADV_I> که<CON>
106 چراکه<CON>
107 سلمنا<CON> که<CON>
108 در<P> حالی<N_SING> که<CON>
109 درست<ADJ> که<CON>
110 همان<DET> طوری<N_SING> که<CON>
111 انگار<CON> که<CON>
112 آنچنان<ADV> که<CON>
113 بگونه‌ای<P> که<CON>
114 بی<P> تو<PRO> که<CON>
115 این<PRO> که<CON>
116 آنچه<PRO> که<CON>
117 هرچه<ADV> که<CON>
118 وقتی<CON> که<CON>
119 هرکه<PRO>
120 هرگاه<CON>
121 هرچه<ADV>
122 چنان‌چه<CON>
123 چنانچه<CON>
124 آنچه<PRO>
125 وقتی<CON>
126 تا<CON>
127 اگرچه<CON>
128 گرچه<CON>
129 اگر<CON>
130 اگه<CON>
131 الا<CON>
132 از<P> بس<ADV>
133 علی‌رغم<CON>
134 علیرغم<CON>
135 هرچند<CON>
136 هروقت<CON>
137 هر<DET> کجا<N_SING>
138 انگار<CON>
139 موقعی<CON>
140 چون<CON>
141 وزانرو<CON>
142 زیرا<CON>
143 ولو<CON>
144 آنگاه<CON>
145 اگرت<CON>
146 گرت<CON>
147 که<PRO>
148 که<CON>
```

Figure 4. Boundary words and their sequence

After finishing processing the sentences containing boundary words, the rest containing two or more verbs can be divided into two groups: main-subordinate complex sentences with complementizers omitted and compound complex sentences. As the relationships between the clauses of compound complex sentences are equal and “<C>” is not needed, the method of Verbal Disconnection should be used to process compound complex sentences. As for compound complex sentences, the coordinate conjunctions should be identified first via the statistically based approach and manually verification. The result is as follows.

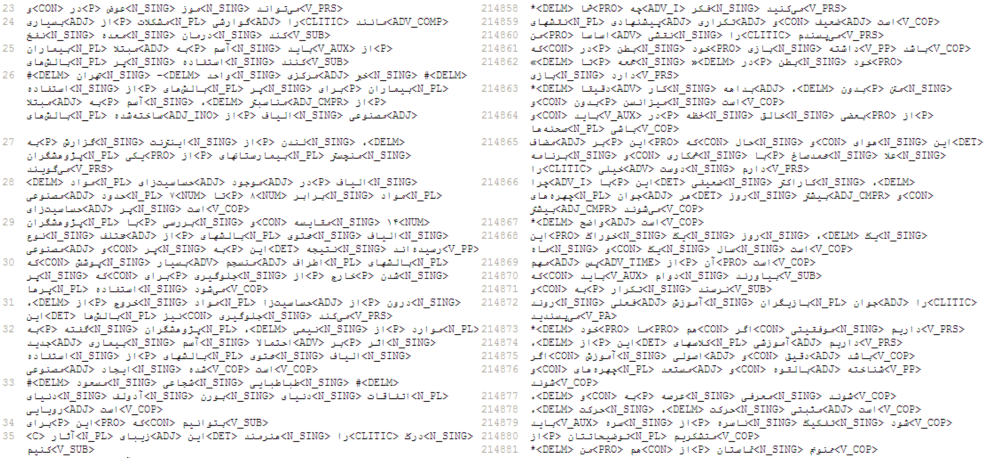
1. فقطنه<CON>…بلکه<CON>, تنهانه<ADV>… بلکه<CON>, هم<CON>…هم<CON>, یا<CON>…یا<CON>, چه<CON>…چه<CON>, نه<CON>…نه<CON>, خواه<CON>…خواه<CON>

2. لیکن<CON>, و<CON>, یا<CON>, اما<CON>, ولی<CON>, والا<CON>, فلذا<CON>, منتهی<CON>, منتها<CON>, وگرنه<CON>, معذلک<CON>, لکن<CON>, بنابراین<CON>, لیک<CON>, لهذا<CON>, معالوصف<CON>, ورنه<CON>, کاندر<CON>, بهسک<CON>, هرآینه<CON>, ولیکن<CON>, و<CON>, امّا<CON>

When a conjunction introduces a compound complex sentence, the word before it is always a verb. For the other sentences, if there is a conjunction together with a verb, the sentence is a compound complex sentence and the method of Verbal Disconnection will be used. The rest containing two or more verbs belong to the main-subordinate complex sentences with complementizers omitted. So, “<C>” will be added when the method of Verbal Disconnection is used to process the sentences.

Design of the System of Extracting Persian Simple Sentences

In this step, the C# language is applied to implement the above algorithm and the number of the simple sentences obtained is 214881, as is shown in Fig. 5,



```
23  و<CON> در<P> عوض<N_SING> موز<N_SING> می‌تواند<V_PRS>
24  بسیاری<ADJ> از<P> مشکلات<N_PL> گوارشی<ADJ> را<CLITIC> مانند<ADV_COMP>
    نفخ<N_SING> معده<N_SING> درمان<N_SING> کند<V_SUB>
25  بیماران<N_PL> مبتلا<ADJ> به<P> آسم<N_SING> باید<V_AUX> از<P>
    بالشهای<N_PL> پر<N_SING> استفاده<N_SING> کنند<V_SUB>
26  #<DELM> تهران<N_SING> -<DELM> واحد<N_SING> مرکزی<ADJ> خبر<N_SING> #<DELM>
    استفاده<N_SING> از<P> بالشهای<N_PL> پر<N_SING> برای<P> بیماران<N_PL>
    مبتلا<ADJ> به<P> آسم<N_SING> ،<DELM> مناسبتر<ADJ_CMPR> از<P>
    بالشهای<N_PL> ساخته‌شده<ADJ_INO> از<P> الیاف<N_SING> مصنوعی<ADJ>

27  به<P> گزارش<N_SING> اینترنت<N_SING> از<P> لندن<N_SING> ،<DELM>
    پژوهشگران<N_PL> یکی<PRO> از<P> بیمارستانهای<N_PL> منچستر<N_SING>
    می‌گویند<V_PRS>
28  <DELM> مواد<N_PL> حساسیت‌زای<ADJ> موجود<ADJ> در<P> الیاف<N_SING>
    مصنوعی<ADJ> حدود<N_PL> ۷<NUM> تا<P> ۸<NUM> برابر<N_SING> مواد<N_PL>
    حساسیت‌زای<ADJ> پر<N_SING> است<V_COP>
29  پژوهشگران<N_PL> با<P> بررسی<N_SING> و<CON> مقایسه<N_SING> ۱۴<NUM>
    نوع<N_SING> مختلف<ADJ> از<P> بالشهای<N_PL> محتوی<N_SING> الیاف<N_SING>
    مصنوعی<ADJ> و<CON> پر<N_SING> به<P> این<DET> نتیجه<N_SING> رسیده‌اند<V_PP>
30  که<CON> پوشش<N_SING> بسیار<ADV> منسجم<ADJ> اطراف<N_PL> بالشهای<N_PL>
    پر<N_SING> که<CON> برای<P> جلوگیری<N_SING> از<P> خارج<P> شدن<N_SING>
    پرها<N_PL> استفاده<N_SING> می‌شود<V_COP>
31  ،<DELM> از<P> خروج<N_SING> مواد<N_PL> حساسیتزا<ADJ> از<P> درون<N_SING>
    این<DET> بالشها<N_PL> نیز<CON> جلوگیری<N_SING> می‌کند<V_PRS>
32  به<P> گفته<N_SING> پژوهشگران<N_PL> ،<DELM> نیمی<N_SING> از<P> موارد<N_PL>
    جدید<ADJ> بیماری<N_SING> آسم<N_SING> احتمالا<ADV> بر<P> اثر<N_SING>
    استفاده<N_SING> از<P> بالشهای<N_PL> محتوی<N_SING> الیاف<N_SING>
    مصنوعی<ADJ> ایجاد<N_SING> شده<V_COP> است<V_COP>
33  #<DELM> مسعود<N_SING> شجاعی<N_SING> طباطبایی<N_SING> #<DELM>
    دنیای<N_SING> آدولف<N_SING> بورن<N_SING> دنیای<N_SING> اتفاقات<N_PL>
    رویایی<ADJ> است<V_COP>
34  برای<P> این<PRO> که<CON> بتوانیم<V_SUB>
35  <C> آثار<N_PL> زیبای<ADJ> این<DET> هنرمند<N_SING> را<CLITIC> درک<N_SING>
    کنیم<V_SUB>
```

```
214858 *<DELM> شما<PRO> چه<ADV_I> فکر<N_SING> می‌کنید<V_PRS>
214859 نقشهای<N_PL> پیشنهادی<ADJ> تکراری<ADJ> و<CON> ضعیف<ADJ> است<V_COP>
214860 من<PRO> اساسا<ADV> نقشی<N_SING> را<CLITIC> می‌پسندم<V_PRS>
214861 که<CON> در<P> بطن<N_SING> خود<PRO> بازی<N_SING> داشته<V_PP> باشد<V_COP>
214862 «<DELM> تا<P> جمعه<N_SING> »<DELM> در<P> بطن<N_SING> خود<PRO>
       بازی<N_SING> دارد<V_PRS>
214863 *<DELM> دقیقا<ADV> کار<N_SING> بداهه<ADJ> ،<DELM> بدون<P> متن<N_SING>
       و<CON> بدون<P> میزانسن<N_SING> است<V_COP>
214864 و<CON> باید<V_AUX> در<P> خطه<N_SING> خالق<N_SING> بعضی<PRO> از<P>
       صحنه‌ها<N_PL> باشی<V_COP>
214865 مضاف<ADJ> بر<P> این<PRO> که<CON> حال<N_SING> و<CON> هوای<N_SING> این<DET>
       برنامه<N_SING> و<CON> همکاری<N_SING> با<P> محمدصالح<N_SING> علا<N_SING>
       را<CLITIC> خیلی<ADV> دوست<N_SING> دارم<V_PRS>
214866 چرا<ADV_I> با<P> این<DET> ضعیفی<N_SING> کاراکتر<N_SING> ،<DELM>
       چهره‌های<N_PL> جوان<ADJ> هر<DET> روز<N_SING> بیشتر<ADJ_CMPR> و<CON>
       بیشتر<ADJ_CMPR> می‌شوند<V_COP>
214867 *<DELM> واضح<ADJ> است<V_COP>
214868 این<PRO> خوراک<N_SING> یک<N_SING> روز<N_SING> ،<DELM> یک<N_SING>
       ماه<N_SING> و<CON> یک<N_SING> سال<N_SING> است<V_COP>
214869 مهم<ADJ> پس<ADV_TIME> از<P> آن<PRO> است<V_COP>
214870 که<CON> باید<V_AUX> دوام<N_SING> بیاورند<V_SUB>
214871 و<CON> به<P> تکرار<N_SING> نرسند<V_SUB>
214872 روند<N_SING> فعلی<ADJ> آموزش<N_SING> بازیگران<N_PL> جوان<ADJ> را<CLITIC>
       می‌پسندید<V_PA>
214873 *<DELM> خود<PRO> ما<PRO> هم<CON> اگر<CON> موفقیتی<N_SING> داریم<V_PRS>
214874 ،<DELM> از<P> این<DET> کلاسهای<N_PL> آموزشی<ADJ> داریم<V_PRS>
214875 اگر<CON> آموزش<N_SING> اصولی<ADJ> و<CON> دقیق<ADJ> باشد<V_COP>
214876 و<CON> چهره‌های<N_PL> مستعد<ADJ> و<CON> بالقوه<ADJ> شناخته<V_PP>
       شوند<V_COP>
214877 ،<DELM> و<CON> به<P> عرصه<N_SING> معرفی<N_SING> شوند<V_COP>
214878 ،<DELM> حرکت<N_SING> ،<DELM> حرکت<N_SING> مثبتی<ADJ> است<V_COP>
214879 باید<V_AUX> سره<N_SING> از<P> ناسره<N_SING> تفکیک<N_SING> شود<V_COP>
214880 از<P> توضیحاتتان<N_PL> متشکریم<V_COP>
214881 *<DELM> من<PRO> هم<CON> از<P> تماستان<N_SING> ممنونم<V_COP>
```

Figure 5. Simple sentences obtained automatically

Conclusion

This paper aims at keeping as many original sentence elements as possible for the extracted simple sentences. The pattern “verb+clause” is very common in Persian, so this paper adds “<C>” in an attempt to mark the clause position. Although the method Verbal Disconnection can process most Persian complex sentences, yet it can’t distinguish main-subordinate complex sentences and compound complex sentences. However, as the clauses of compound complex sentences have an equal position, “<C>” is not needed. According to the method Verbal Disconnection, as long as there is a coordinate conjunction in a sentence, the sentence should be disconnected from behind the verb without “<C>” and thus the position of the clause can’t be determined. Therefore, Verbal Disconnection is not suitable for multi-complex sentences, which contain both main-subordinate complex sentences and compound complex sentences. In the actual text, multi-complex sentences are very common. Therefore, the method of Boundary Word Interception is also proposed for processing the main-subordinate complex sentences.

The sequence among the steps of the algorithm of extracting Persian simple sentences above is determined through repeated experiments and on the basis of Persian grammar rules. Therefore, almost all the complex sentences in UPC have been processed with one exception that the complementizer of main-subordinate complex sentences is omitted and coordinate conjunctions exist in very few multi-complex sentences. This leads to the loss of “<C>” after the verb. As for UPC, this is a rare case and the correction is done manually, but if the algorithm in this paper is applied to process other Persian corpora and the case above is not rare, the approach of counting occurrence frequency can be used to correct the rare pattern of a verb.

To sum up, good experimental results have proved the effectiveness of the algorithm proposed in this paper. As all the simple sentences are obtained from the authentic corpus, they can also be important resources for Persian linguistic research.

References

[1] Mahmood, B., The Role of the Corpus in Writing a Grammar: An Introduction to a Software. Iranian Journal of Linguistics, (19), 2004.

[2] Mojgan, S., Morphosyntactic Corpora and Tools for Persian. Doctoral dissertation, Uppsala University. Studia Linguistica Upsaliensia 16, 2015.

[3] Muhtar, M., Guli, E. & Abdura, J., Research on the Recognition of Modern Uyghur Simple Sentences. Computer CD Software and Applications, (13), pp. 211-212, 2014. (In Chinese)

[4] Li, Y. C., Feng, W. H., Zhou, G. D., et al. Research on Chinese Clause Identification Based on Comma. Acta Scientiarum Naturalium Universitatis Pekinensis, (1), pp. 1-8, 2013. (In Chinese)

[5] Jiang, Y. R. & Song, R., Topic Clause Identification Based on Generalized Topic Theory. Journal of Chinese Information Processing, 26(5), pp. 114-119, 2012. (In Chinese)

[6] Ma, F., Wu, B. M. & Wang, B. X., Clause Recognition Method for English-Chinese Machine Translation. Journal of Information Engineering University, 7(2), pp. 193-196, 2006. (In Chinese)

[7] Molina, A. & Pla, F., Clause Detection using HMM. In Proceedings of the CoNLL-2001. Toulouse, France, pp. 70-72, 2001.

[8] Erik, F. & Tjong, K. S., Memory-Based Clause Identification. In Proceedings of the CoNLL-2001. Toulouse, France, pp. 67-69, 2001.

[9] Available at: http://stp.lingfil.uu.se/~mojgan/UPC.html

[10] Zhang, L.M., Persian Grammar. Guangzhou: World Book Publishing, pp. 157-272, 2016.(In Chinese)