Persian Discourse Treebank and coreference corpus

Azadeh Mirzaei, Pegah Safari
Department of Linguistics, Allameh Tabataba'i University, Tehran, Iran; Department of Technology, Alzahra University, Tehran, Iran
azadeh.mirzaei@atu.ac.ir, phsafari@yahoo.com

Abstract
This research addresses the investigation of intra-document relations based on two major approaches: discourse analysis and coreference resolution which results in building the first Persian discourse Treebank and a comprehensive Persian coreference corpus. In discourse analysis, we have explored sentence-level relations defined between clauses in complex sentences. So we specified 34682 discourse relations, the sense of the relations, their arguments and their attributes mainly consisted of the source of the message and its type. Our discourse analysis is based on a corpus consisted of 30000 individual sentences with morphologic, syntactic and semantic labels and nearly half a million tokens. Also 18336 of these sentences are double-annotated. For coreference annotation, since a document-based corpus was needed, we prepared a new corpus consisted of 547 documents and 212646 tokens which is still under development. We enriched it with morphological and syntactical labels and added coreference information at the top. Currently, we have annotated 6511 coreference chains and 21303 mentions with a comprehensive annotation scheme to compensate some specification of Persian such as being pro-drop or lacking gender agreement information.

Keywords: Persian discourse Treebank, Persian Coreference Corpus, logical relation, mention, referent.

1. Introduction

In this research we addressed the investigation of intra-document relations with two major approaches: discourse analysis and co-reference resolution. In discourse analysis, we can inspect the logical relations between sentences inside a document and also we can explore sentence-level relations defined between clauses in complex sentences. Since there was already a rich annotated Persian corpus consisted of 30000 individual sentences, we augmented sentence-level discourse information on the top of its morphologic, syntactic and semantic layers. But for coreference annotation, a document-based corpus was needed; because in this annotation, the relation between a referent and its mention is defined inside a document and it is not anymore restricted to an individual sentence. So we prepared a new corpus and enriched it with morphological and syntactical labels to be used for learning processes in the future. Then we added coreference information at the top. Currently, we are adding a new document-level discourse annotation to our new corpus as well. So in this paper, first of all, we describe the process of preparing the first Persian discourse Treebank and then elaborate the preparation process of our first comprehensive Persian coreference corpus.

2. Persian Discourse Treebank

Persian Discourse Treebank (PerDTB) as the first discourse corpus in Persian, has been developed based on the schema of Penn Discourse Treebank (weber et al., 2003; Prasad et al., 2008) which has been used in discourse projects of the other languages such as Arabic (Al-Saif and Markert, 2010), Chinese (Zhou and Xue, 2012), Czech (Mladov´a et al., 2008), Hindi (Oza et al., 2009), Italian (Tonelli et al., 2010) and Turkish (Zeyrek and Webber, 2008).
The corpus is based on nearly 30,000 sentences which has received morphologic and syntactic labels (Rasooli et al. 2013) and also went through the semantic role labeling process (Mirzaei & Moloodi 2016). Although the corpus is based on individual sentences, it's richly annotated in different levels and it can provide us valuable features for building learners in future. Also with this corpus, we can focus on intra-sentential relations (the relations inside a sentence) which are one of the main types of discourse relations. The corpus consists of 18336 complex sentences and 11646 simple sentences. In annotation process of individual sentences, if the sentence is a complex clause/sentence, according to the systemic functional grammar (Halliday & Matthiessen 2013), we have to specify the logical relations between its clauses, the type of the relations and their attributes while for simple clauses just the type of the relation is specified. In complex sentences where there is no logical relation, if there is a clause showing the source of the message, the attribute of the sentence is annotated as well. In The following examples, the first one shows these kinds of sentences with its attribute specified while in the second one with intra-sentential relation, the connective is marked.

- او با خودش گفت که هر روز دو برابر روز قبل گل‌کاری خواهم کرد.

He said to himself that he would plant each day twice the previous day. (complex clause without any relation)

- چون لکهٔ تمشک مقاوم است باید محل لکه را در محلول پودر رختشویی بخیسانید و اگر شستید و نرفت باید مقداری دوغ یا ماست رویش بریزید.

As the stain of blueberry is resisted, you have to steep it inside the detergent and if you have washed it and it still remained, you have to pour yoghurt on it. (complex clause with logical relation)

Attribute of the relation is defined according to the PDTB standards and it mainly consists of source and type. By source we mean the source of the message which can be the writer of the sentence (Wr), other person mentioned in the text (Ot) or one arbitrary person not mentioned (Arb). In our corpus there are 30795 writer (Wr), 5545 other (Ot) and 552 arbitrary (Arb) sources. Type refers to the objectivity or subjectivity of the message/ sentence stated by source. It can show us whether it is an assertion, declarative sentence or just a subjective sentence to show the source's attitudes. So the type is categorized into four groups: assertion, facts, beliefs and eventualities. Assertion and belief are both similar as they force the agent/source to be committed to the truth of the sentence while they are different in the commitment degree. Eventualities are completely different and they show the intention or attitude of the source. In our corpus, type is annotated based on the category of its verb and it is classified into four main groups which is the same as PDTB: Communicative verbs (Comm) for assertions,

Propositional Attitude verbs (PAtt) for beliefs, Factive or semi-factive verbs (Ftv) for facts and Control verbs (Ctrl) for eventualities. In our Treebank there are 2885 Comm, 561 PAtt, 1019 Ftv and 3304 Ctrl.

3. Annotation Procedure of PerDTB

From linguistic point of view, Persian Discourse Treebank is based on the systemic functional grammar and logical metafunction concept (Halliday & Matthiessen 2013) and its annotation scheme is based on the standards used in the Penn Discourse Treebank (weber et al., 2003; Prasad et al., 2008). In this scheme, first of all, the logical relations between clauses are specified. If there is any connective, the relation is categorized into explicit and the sense of the connective is determined. Since the sense classification is one of the main procedures of annotation, it has been elaborately described in the next section. After that, the arguments of the relation should be specified. Persian is a free-word-order language, so the position of connective can be sometimes permutated in the sentence and also the arguments can take different positions (one can precede the other or they can be nested). When there is no connective, the annotator inserts a meaningful connective which is added to the discourse information layer, specifies its sense and classifies the relation as implicit. AltLex is another discourse relation type which is used when the relation is alternatively lexicalized by some non-connective expression in the sentence and the last relation type, EntRel (entity-based relation), is specified when one clause contains an entity and the other one describes it. Also PDTB contains another relation type, NoRel, which is not defined in our Treebank due to the annotation of individual sentences. Table1 shows the frequency of relation types in our corpus. The last row shows simple clauses or complex sentences without logical relation in which just the attribution is specified and in the first row, Explicit and AltLex relations are counted as joint. Its reason has been elaborated in Annotators Agreement section.

Relation Types	#frequency
Explicit+AltLex	13129
Implicit	1108
EntRel	54
Clause	20371

Table 1: Distribution of relation types in PerDTB

At the end, for all of the relation types except EntRel, we can specify the attributes of the relation or its individual arguments which can take different forms. For example, to express the source of a relation we can use a clause, a prepositional phrase, an adjective phrase, etc. (e.g., scientists say, according to scientists, quoted by scientists). Also in order to facilitate the annotation process, we have developed an annotating application with administrative panels to supervise the process and guide the annotators if it was needed. Figure 1 shows the general view of the program.

4. Sense Annotation

In Explicit, Implicit and AltLex relations, sense annotation of the connective or the lexical structure is one of the main steps in the process. According to PTDB, sense has four main classes in the first level (temporal, expansion, contingency and comparison) and each class is again divided into sub-categories as the second level and even more elaborately sub-divided into sub-types in the third level. Table 2 shows the hierarchical structure of the senses.

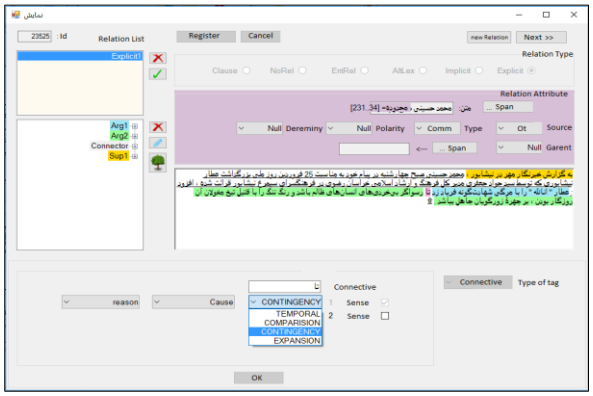

Figure 1: The general view of the annotation tool

Classes (first level)	Types (second level)	Sub-types (third level)
Comparison	Contrast, Pragmatic contrast, Concession, Pragmatic Concession	Juxtaposition, opposition (contrast) Expectation, contra-expectation (concession)
Contingency	Cause, Pragmatic cause, Condition, Pragmatic Condition	Reason, result (cause) Justification (pragmatic cause) Hypothetical, general, unreal present, unreal past, factual present, factual past (condition) Relevance, implicit assertion (pragmatic condition)
Expansion	Conjunction, Instantiation, Restatement, Alternative, Exception, List	Specification, equivalence, generalization (restatement) Conjunctive, disjunctive, chosen alternative (alternative)
Temporal	Asynchronous, Synchronous	Precedence, succession (asynchronous)

Table 2: The hierarchical structure of the senses

It is noteworthy that most of connectives are homonym and as a result they have different senses in different contexts such as “and”/و/ or “that”/که/ which sense can be categorized into all of the four main classes based on their context. Also there would be multiple sub-categories for an individual connective like “as”/چون/ (condition and cause). This homonymy can lead to different perception of the text and it can be the main reason of disagreement between annotators. The following examples show this homonymy and different sense classes for a single connective:

- پارسا هم چون حوصله‌اش سررفت شروع به لجبازی کرد. (contingency, cause)
As Parsa was bored, he started to become obstinate.

- چون تو با منی قلمرو را با یزدان دو بخش خواهم کرد. (contingency, condition)

As you are with me, I would divide the realm with God.

- چون او وارد شد، همه بر پای خاستند. (temporal)

As he entered, everyone stood up.

In the next section, we have presented some of the useful statistics of sense distribution in our Treebank which can lead to a better perception of the structure.

5. Statistics of PerDTB

PerDTB consists 29982 individual sentences. Nearly 61% of them are complex sentences with more than one verb and just about 38% are simple clauses. These complex clauses are not guaranteed to have logical intra-relations which can be proved by 40% of complex sentences classified as Clause. In most of them, there is a clause playing the role of attribute and showing the source of the message. Table 3 gives these statistics about the corpus.

#Sentences	29982
Average Length of Sentences	16.61
#Verbs	62889
# Sentences with One Verb	11617
# Sentences with Two Verbs	9917
# Sentences with More than Two Verb	8419

Table 3: Statistics of PerDTB

Table 4 shows the distribution of sense classes in three discourse relation types. Expansion is the most dominant sense with about 41% which shows that an individual sentence is more used for Expansion while Comparison with the least frequency (nearly 9%) shows that it is usually expressed through multiple sentences. This restriction to one or multiple sentences can be considered as a feature for sense classification.

Class	Explicit+AltLex	Implicit	Total
Temporal	1813	47	1860
Contingency	5383	385	5768
Comparison	1330	75	1405
Expansion	5644	673	6317
Total	14170	1180	15350

Table 4: Distribution of sense classes in PerDTB

Table 5 and 6 show the top five tags of the second and the third level of sense in the corpus.

Second level of Sense (types)	#frequency
Conjunction	4269
Cause	3548
Condition	2195
Restatement	1385
Synchronous	1014

Table 5: Top five types of senses (in the second level)

third level of Sense (sub-types)	#frequency
specification	1149
reason	1498
equivalence	196
juxtaposition	410
generalization	40

Table 6: Top five sub-types of senses (in the third level)

As we have mentioned, due to the homonymy, the sense classification of some connectives can be so confusing. Table 7 lists the six top homonym connectives.

Connective	Senses
در حالی که (during)	Temporal (30), Comparison (22), Contingency (2)
بلکه (but)	Comparison (69), Contingency (2), Expansion (29)
وقتی (when)	Temporal (295), Comparison (1), Contingency (133), Expansion (1)
تا (until)	Temporal (69), Contingency (762), Expansion (2)
که (that)	Temporal (208), Comparison (48), Contingency (800), Expansion (327)
و (and)	Temporal (360), Comparison (287), Contingency (925), Expansion (4591)

Table 7: Top six homonym connectives

Table 8 shows the top ten connectives which length is just one-token while table 9 shows the top ten connectives with more than one token in length. These connectives are compound or they are prepositional/ noun phrases that always appear continuously and their function is the same as one-token connectives.

Connective	#frequency	Connective	#frequency
و (and)	5993	وقتی (when)	377
اگر (if)	1612	ولی (however)	253
که (that)	1338	یا (or)	242
تا (until)	846	چون (since)	220
اما (but)	395	زیرا (because)	151

Table 8: Top ten one-token connectives

Connective	count	Connective	count
هنگامی که (when)	93	وقتی که (when)	53
در حالی که (during)	91	برای این که (because)	40
زمانی که (when)	67	در نتیجه (As a result)	38
هر گاه (when)	64	چرا که (because)	33
در صورتی که (if)	60	تا زمانی که (until)	25

Table 9: Top ten multi-token connectives

Table 10 shows the top ten discontinuous connectives which are consisted of two or multiple parts. Each part shows one piece of the logical relation and all of the parts, together, show one relation.

Connective	count	Connective	count
نه تنها | بلکه (Not only| But also)	30	گرچه | اما (Although| but)	9
نه | نه (Neither| Nor)	29	اگر | پس (If| Then)	9
هم | هم (Also| Also)	26	چه | چه (Either| Either)	7
یا | یا (Either| Or)	21	هرچند | اما (Although| But)	7
اگرچه | اما (Although| But)	12	اگر | دراین صورت (If| Then)	6

Table 10: Top ten discontinuous connectives

6. Inter-annotation Agreement of PerDTB

Our PerDTB annotation group consisted of four PhD candidates in linguistics and one MA graduated of Persian

language and literature. They were native Persian speakers and were presented a comprehensive guideline, describing all the logical relations with abundant examples. In order to measure their inter-annotation agreement, in the next phase, we double annotated 18336 sentences. We have used the kappa statistics (Cohen, 1960) for our measurement purpose which is defined with respect to the probability of inter-annotator agreement, P(A), and the agreement expected by chance, P(E):

$$k = \frac{\mathrm{P(A)} - \mathrm{P(E)}}{1 - \mathrm{P(E)}}$$

Our results show that the inter-annotator agreement of sense classification is 'good' with kappa value of 0.769. Also the agreement of Implicit, EntRel and Clause relation is 'very good' with k=0.856 but the agreement of Explicit and AltLex decreases to 'moderate' level with k=0.446. The reason of this disagreement can be interpreted as the state of the language. Persian, typologically is in a transitional state and it is moving from an agglutinative language toward becoming more analytic. The disagreement shows this transition. It shows that although the annotators agree on the logical function of a structure but they don't have much agreement on its essence (connective in Explicit vs. non connective expression in AltLex). This situation generally doesn't affect the annotation of logical relations between clauses.

7. Persian Coreference Corpus (PerCoref)

After developing our first Persian discourse corpus and inspecting logical relations inside complex sentences, we put a step forward to investigate intra-document relations through coreference resolution. Coreference resolution is one of the building blocks for high level NLP tasks such as question-answering systems, summarization, machine translation, etc. these reasons provoked preparation of coreference corpus in different languages like English (Hirshman, 1998), Dutch (Hendrickx, et al., 2008), Japanese (Iida, et al., 2007), Polish (Ogrodniczuk, et al., 2013), Spanish and Catalan (Recasens, & Martí, 2010). Also in Persian different coreference resources were prepared but as far as we know there is not yet any comprehensive coreference corpus in Persian to compensate some features of the language. For example, Persian is a pro-drop language, there is no gender agreement and sometimes the number agreement is ignored. So developing a comprehensive coreference corpus to cover all kinds of referential expressions and including complementary information to cover the eliminated info was needed. We tried to include any valuable information to compensate these restrictions in our corpus with elaborated coreference relation tags and covering different referential candidates. The tagset of the project is based on a coreference guideline by Komen (2009) and the theoretical issues about cohesive strategies in Halliday and Hasan (2014).

For our purpose, the previously used corpus for PerDTB was based on individual sentences and it was not suitable enough to cover coreference relations inside a cohesive text. So we developed a document based corpus. At the moment it has overall 212646 tokens and 547 documents which are majorly crawled from news articles and have been processed manually to receive morphological and syntactical labels.

8. Annotation Procedure of PerCoref

As the annotation scheme of our corpus, first of all, we specify any referential candidate (mention) which can be pronoun, noun phrase or verb for null-subject sentences. After that, their references are specified and the type of each reference is marked. Generally, the reference type can be categorized into three groups: direct reference, indirect reference or no-reference. Direct reference is subdivided into anaphoric (appear before mention) or cataphoric (appear after mention) while indirect reference as it indirectly refers to the speaker or listener of the text, is classified into speaker reference (refereed to speaker) or addressee reference (refereed to listener). However, some pronouns such as "nobody"/هیچکس/ or "other"/دیگران/ actually doesn't refer to any specific entity and we have marked them as no-reference pronouns. The following examples elaborate each type of references.

- علی به خانه آمد. همه از آمدن او خوشحال بودند. (Anaphoric)

Ali came home. Everyone was pleased with his return.

- من دیروز او، یعنی علی، را دیدم. (Cataphoric)

I see him, I mean Ali, yesterday.

- ما باید نسبت به جامعه خود بیشتر مسئول باشیم. (Indirect reference)

We have to be more responsible for our society.

- هنوز هیچکس علت را نمی‌داند. (No-reference)

No one knows the reason yet.

Table 11 shows the distribution of reference types in our corpus.

Reference Types		#frequency
Direct Reference	Anaphoric	18713
	Cataphoric	1485
Indirect Reference		281
No-Reference		824
Total		21303

Table 11: Distribution of reference types in PerCoref

In the next step, the type of the relation between the mention and its reference is investigated. This relation type is defined for Direct and Indirect references while no-reference expressions just receive some complimentary information including the number (single, plural) and a label to show the semantic class of the mention consisted of: personal, time, place or other. In the next section the coreference relation types are completely described. For the span selection of mentions, only the head of the phrasal group is included and also for each mention its closest referent is annotated. However, sometimes the referent is discontinuous with multi parts. In these cases, we have specified all separated parts in the referent span. At present, our corpus contains 6511 coreference chains and 21303 mentions. By chains, we mean the mentions refereed to the same entity.

For annotation procedure, we have developed an annotator program suited to our specific scheme. Figure 2 shows the overall view of the application. Also our annotation group consisted of one linguistic PhD, five linguistic PhD candidates, one MA graduated in Persian language and literature and one BA of English translation who were native Persian speakers supplied with a comprehensive guideline and adequate examples.

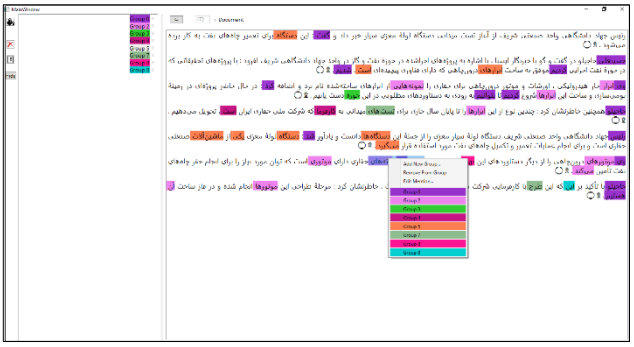

Figure 2: The overall view of the annotation tool

9. Coreferece Relation Types

Coreference relation type is one of the most valuable information added to our corpus during the annotation process. It shows the relationship between the mention and its referent. Direct and Indirect referents have their own coreference relation type categories. If the type of the referent is indirect, it is speech reference and the mention is first-person plural pronoun then there are three relation types: Exclusive-we which is defined when the mention refers to the speaker and his/her clique. Co-present-we in which the mention refers to the speaker and all of the persons who are listening to his/her notes and as the last type, all-inclusive-we which doesn't refer to any specific person and just indicates 'we' as human being. For the first-person single pronoun, all-inclusive-I is the same as all-inclusive-we but deictic-I simply indicates the writer or speaker. When the type of the referent is addressee reference, there are two major relation types: deictic and generic; the former refers to a specific person and the latter doesn't indicate any specific person but the human being as a whole. As in Persian there are two pronouns for second-person: single and plural, the type is classified into four groups. On the other hand, for direct referent, there are six major groups: identity, inferred, quantifier, cross-speech, event and person/number suffix on verb. Identity is defined when the mention and its referent both indicate the same entity in the real world. Since there is number agreement in this case, single or plural is specified. Inferred relation type unlike identity happens when the mention and referent are not identical but the first one infers the second in some way. According to this inferring style, the relation type is subdivided into five categories: hyponymy, meronymy, member-collection, antonymy and portion-mass. In hyponymy, the mention is a type of its referent and the mentions inside a chain are their co-hyponyms. In meronymy, the mention is not a type of referent but it is one of its parts. Again the mentions are co-meronyms with each other in a chain. Member-collection refers to the situation when all of the mentions in a chain can be linked together as members of a larger collection like trees in a jungle. Antonymy as its name suggests is used when the mentions are their antonyms. In portion-mass category, the mention is a measurable portion of referent as the mass. Quantifier as the third main relation type is defined to support immeasurable portions of referent used as mention like "some"/تعدادی/ or "few"/برخی/. Cross-speech, the fourth type, is used when we have direct speech in the text. In this situation the relation between mention in direct speech and the referent in indirect one is marked as cross-speech to register the disagreement in the pair. In event type, mention refers to an event introduced in the text and we annotate the verb of the clause as its referent. The last relation type is person/number suffix on verb which shows the subject of the null-subject sentences. Since Persian is a pro-drop language, including this relation type would provide valuable information. Also the number and the person of dropped pronoun are specified. The examples below show the main coreference relation types annotated in our corpus:

- من دیروز علی را دیدم و با او حرف زدم. (Identity)

I saw Ali yesterday and talked to him.

- علی به زمین افتاد. دستش صدمه زیادی دید. انگشتش شکست. مچش دررفت و زانویش کبود شد. (Inferred-hyponymy)

Ali fell down. His hand was hurt badly. His finger was broken, his wrist was displaced and his elbow was bruised.

- دیروز به مغازه رفتم. چشمم به شکر افتاد و دو کیلو خریدم. از آنجا که میوه خوبی هم داشت، مقداری میوه هم خریداری کردم.(Inferred-Portion-mass, Quantifier)

Yesterday, I went to the grocery. As I saw sugar, I bought two kilos. Since the fruits were fresh, I bought some fruits as well.

- آنها مشقشان را انجام ندادند. این خیلی بد است. (Event)

They didn't do their homework. It is so unpleasant.

- امروز از آنها تشکر کردم. دیروز هدیه‌ای به من داده‌بودند. (person/number suffix on verb)

Today I thanked them. Yesterday they had given me a gift.

- دوستم به من زنگ زد و گفت: "تو را دیروز ندیدم. حالت خوب است؟" (Cross-speech)

My Friend called me and asked: "Yesterday I didn't see you. Are you OK?"

10. Statistics of PerCoref

PerCoref consists of 547 documents which have been mainly selected from news articles with 212646 tokens. The corpus is under development and it would become larger. Table 12 shows some useful statistics of our corpus.

Number of documents	547
Number of tokens	212646
Number of sentences	6688
Number of paragraphs	4449
Avg. paragraphs per documents	8.13
Avg. sentences per paragraph	1.5
Avg. tokens per sentence	31.79

Table 12: Statistics of PerCoref

Although the Average number of sentences in paragraphs doesn't sound reasonable but as the last row shows, the sentences are long which means that the number of complex sentences tends to be much more than simple clauses and it would compensate the short average length of paragraphs.

Size of chain	frequency	% of all mentions
1	3692	17.33
2	929	4.36
3	460	2.15
4	287	1.34
5	211	0.99
6-10	521	2.44
11-50	401	1.88
More than 50	11	0.05

Table 13: Length of chains

In PerCoref, we have specified 6511 chains with average size of 3.27 mentions while the longest chain contains 87 mentions. Table 13 shows the frequency of different sizes for the coreference chains.
Table 14 and 15 show the distribution of coreference relation types for direct and indirect references respectively. As we can see, in table 14, Identity with nearly 58% is the dominant relation type for direct reference while cross-speech with less than 5% is the least frequent relation in the corpus. Person/number suffix on verb with significant frequency of about 14% as the second dominant relation shows that Persian speakers are considerably inclined to drop pronouns in their written texts.

Coreference relation type	count	% of count
Identity	11812	58.4
person/number suffix on verb	3021	14.9
Inferred	2005	9.9
Event	1509	7.4
Quantifier	968	4.7
Cross-speech	901	4.4

Table 14: frequency of relations types for direct references

Relation types for indirect referent	count	% of count
Exclusive-we	136	48.3
Co-present-we	75	26.6
Generic-you (plural)	25	8.8
Generic-we	15	5.3
Deictic-you (plural)	12	4.2
Generic-you (single)	9	3.2
Deictic-you (single)	4	1.4
Deictic-I	3	1
Generic-I	2	0.7

Table 15: frequency of relations types for indirect references

In table 15, we can see that excusive-we with 48% and co-present-we with 26% are the most frequent relation types but the usage of the other types is trivial in comparison with these two categories except Generic-you in its plural form with nearly 9% frequency.
Table 16 shows the frequency of sub types for inferred relation type.

Inferred sub type	Count	% of count
hyponymy	581	29.2
Co-hyponymy	343	17.2
Meronymy	387	19.4
Co-meronymy	118	5.9
Member-collection	470	23.6
Portion-mass	60	3
Antonym	14	0.7
Other	14	0.7

Table 16: frequency of inferred sub types

Hyponymy with 29% is the dominant sub type while its pair, co-hyponymy is in the fourth place with 17%. It shows that the length of the hyponym chains is less than their frequency. This situation is so considerable for meronymy and co-meronymy. Meronymy with 19% frequency, as the third frequent sub type, shows that this category occurs significantly while the length of these chains is too short which is shown by co-meronymy with just 5% frequency. Other in the last row, happens when the annotators cannot find suitable sub type for inferred relations.

11. Conclusion

In this paper, first of all with a focus on intra-document relations, we presented the structure and preparation process of the first Persian Discourse Treebank (PerDTB). It is based on the previously morphological and syntactical annotated corpus with nearly 30000 individual sentences concentrated on intra sentential relations. The next version of this Treebank is under development which is based on the documents in coreference corpus with half a million words to cover the other kind of relations such as the relations between two adjacent sentences. After that, in order to explore more intra-document relations, we developed the first comprehensive Persian Coreference Corpus (PerCoref) and described its preparation process. It consists of 547 documents with 212646 tokens which received POS, syntactic and coreference tags and it is still under developed. At the moment, we have annotated 6511 coreference chains and 21303 mentions and we are planning to double annotate our corpus to measure inter-annotator agreement. For more information about releasing notes of this version of PerDTB or the current version of PerCoref, you can visit Peykaregan website (http://www.peykaregan.com) and also for complementary information on PerDTB the following website is available: http://opensourceiran.ito.gov.ir/web/guest/-2.

12. Acknowledgements

Our first corpus, PerDTB, was funded by Iran Information Technology Organization (ITO) and Computer Research Center of Islamic Sciences (CRCIS) and our second corpus, PerCoref, was funded by CRCIS. During this project we had different annotation phases: PerDTB annotation, POS and syntactical annotation as preparation of coreference corpus and annotating PerCoref. We really appreciate the linguists who helped us with these annotation phases: Farzaneh Bakhtiyari (coreference preparation), Parinaz Dadras (all phases), Saeedeh Ghadrdoost-Nakhchi (all phases), Manoucher Kouhestani (PerDTB annotation), Neda Poormorteza-Khameneh (all phases), Mostafa Mahdavi (coreference preparation), Samira Mirzaei (coreference preparation) and Salimeh Zamani (all phases); and other colleagues especially Mahdi Behniafar.

13. References

Al-Saif, A. and Markert, K. (2010). The Leeds Arabic Discourse Treebank: Annotating Discourse Connectives for Arabic. *In Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)*, Valletta, Malta.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychosocial measurement*, 20(1), pp. 37-46.

Halliday, M., Matthiessen, C. M., & Matthiessen, C. (2014). *An introduction to functional grammar*. Routledge.

Halliday, M. A. K., & Hasan, R. (2014). *Cohesion in english*. Routledge.

Hendrickx, I., Bouma, G., Daelemans, W., Hoste, V., Kloosterman, G., marie Mineur, A., Van, J., Vloet, D., and Verschelde, J.-L. (2008). A Coreference Corpus and Resolution System for Dutch. *In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, pp 144–149, Marrakech, Morocco. European Language Resources Association (ELRA).

Hirshman, L. and Chinchor, N. (1998). MUC-7 coreference task definition. Version 3.0. *In Proceedings of the Seventh Message Understanding Conference (MUC-7).*

Iida, R., Komachi, M., Inui, K., and Matsumoto, Y. (2007). Annotating a Japanese Text Corpus with Predicate-Argument and Coreference Relations. *In Proceedings of the Linguistic Annotation Workshop (LAW 2007)*, pp 132–139, Stroudsburg, PA, USA. Association for Computational Linguistics.

Komen, E. R., (2009). Coreference Annotation Guidelines, *http://repository.ubn.ru.nl/bitstream/handle/2066/78810/78810.pdf*.

Mirzaei, A. & Moloodi, A. (2016). Persian proposition Bank. *In Proceedings of the 10th International Language Resources and Evaluation*, Portorož (Slovenia), May, pp. 3828-3835.

Mladová, L., Zikánová, Š., and Hajičová, E., (2008). From Sentence to Discourse: Building an Annotation Scheme for Discourse Based on Prague Dependency Treebank. *In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Oza, U., Prasad, R., Kolachina, S., Meena, S., Sharma, D.M., and Joshi, A., (2009a). Experiments with Annotating Discourse Relations in the Hindi Discourse Relation Bank. *In Proceedings of the 7th International Conference on Natural Language Processing (ICON-2009)*, Hyderabad, India.

Ogrodniczuk, M., Głowi´nska, K., Kope´c, M., Savary, A., and Zawisławska, M. (2013). Polish Coreference Corpus. *In Vetulani, Z., editor, Proceedings of the 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pp 494–498, Pozna´n, Poland.

Prasad, R., Dinesh, N., Lee A., Miltsakaki E., Robaldo L., Joshi A. and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco, May.

Rasooli, M. S., Kouhestani, M. and Moloodi, A. (2013). Development of a Persian syntactic dependency treebank. *In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* Atlanta, USA, pp. 306-314.

Recasens, M. and Martí, M. A. (2010). AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44(4):315–345.

Tonelli, S., Riccardi, G., Prasad, R., and Joshi, A. (2010). Annotation of Discourse Relations for Conversational Spoken Dialogs. *In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.

Webber, B., Joshi A., Stone M., and Knott, A. (2003). Anaphora and discourse structure. Computational Linguistics, 29(4):545–587.

Zeyrek, D. and Webber, B., (2008). A Discourse Resource for Turkish: Annotating Discourse Connectives in the METU Corpus. *In Proceedings of IJCNLP-2008.* Hyderabad, India.

Zhou, Y. and Xue, N. (2012). PDTB-style Discourse Annotation of Chinese Text. *In Proc. 50th Annual Meeting of the ACL*, pp: 60- 77, Jeju Island, Korea.