

# A Corpus-based Study on Abstract Anaphora Resolution

Ping Yu

Department of Linguistics  
University of Michigan  
[yup@umich.edu](mailto:yup@umich.edu)

## Abstract

This paper is a corpus-based study on abstract anaphora resolution, especially on classification of anaphors. To get a consistent analysis of abstract anaphors, I propose that NPs denoting abstract entities are possible abstract referents besides other linguistic units claimed in the literature. I claim that the forms of anaphors, especially types of nouns for *this N* and *that N* are the first criteria in anaphor classification. Following the previous study, I assume that the predicate information is the second criteria for abstract anaphor classification. I manually labeled a small size of sample corpora and found some preliminary results on these two criteria.

## 1 Introduction

Anaphor resolution has always been one of the challenging problems in computational linguistics. A considerable range of research has been done.<sup>1</sup> However, most approaches developed have been restricted to deal with anaphoric relations between nominal pronominal anaphors and NP-antecedents with reference to individual entities. This widespread trend is taken as the representative of anaphora resolution as a whole in the study of anaphora resolution. On the contrary, there are very few approaches done on anaphoric relations

between anaphors and higher order entities, such as situations, facts, events, propositions, states, actions, etc. Compared to the developed stage of individual anaphora resolution, abstract anaphora is still at its infant stage due to the complexness of abstract anaphora. No complete theory has emerged given the complexity of the problem and implementation in abstract anaphora is restricted. Therefore, it is wise to do some empirical study on abstract anaphora resolution. The paper is a corpus-based study on anaphora resolution, especially on the classification of anaphors, which is the first and important step in abstract anaphor resolution. This paper is a preliminary study by using a small size of sample data. The corpus used in this paper is ACE-2.0.<sup>2</sup> I only analyzed the train part of broadcast news. There are 51 news articles, and 107,578 words.

## 2 Forms of Abstract entities

In the paper, I make two assumptions. First, there are some abstract anaphors whose antecedents are previous text spans, and it is possible to resolve their referents through linguistic knowledge. I will not deal with those abstract anaphors whose referents are either not embedded in previous text or need to be inferred beyond linguistic knowledge, such as world knowledge. Secondly, text spans that can be referred to as abstract referents are taggable and can be possibly extracted automatically in implementation.

---

<sup>1</sup> See Mitkov's (2002) review.

---

<sup>2</sup><http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003T11>

## 2.1 Forms of abstract anaphors

Both theoretical studies (Webber, 1991; Asher, 1993) and implementations (Byron, 2002; Eckert and Strube, 1999) of abstract anaphora resolution consider the three over pronouns: *this*, *that* and *it* as possible abstract anaphors when they function as heads by themselves. In this paper, I propose that *this* and *that* are abstract anaphors when they function not only as heads but also as modifiers of noun phrases based on the following observations:

First, *this* and *this N*, *that* and *that N* exhibit similar features. They are sometimes interchangeable, and *this N/that N* provide more specific information in abstract anaphora resolution. Consider (1)<sup>3</sup>.

- (1) Leaders in the African-American community believe the rapid spread of AIDS among blacks has created a national health emergency. What is most striking about **this whole situation** is that every hour seven Americans become infected with HIV.

In (1), *this whole situation* can be substituted by *this*, and *this whole situation* gives a clear hint that it expresses a situation.

Secondly, *this/that N* exhibit restrictions on semantic compatibility with their referents or predicates in the similar way as *this/that* do, such as (2).<sup>4</sup>

- (2) a Fred believes [that John viciously kicked Mary]. This is true.  
b. # Fred believes [that John viciously kicked Mary]. This indicates that they are not getting along at all well.  
c. # Fred believes [that John viciously kicked Mary]. This fact indicates that they are not getting along at all well.

*That*-clause as the complement of *believe* denotes a proposition. (2a) is good because the propositional presupposition of the predicate verb has been satisfied. (2b) and (2c) are not good. The predicate denotes both *this* and *this fact* have factive readings,

but the presupposition of the factive readings is not fulfilled in the previous sentence.

Theoretically, Gundel et al. (1993) argue that *this N*, *that N* have similar referential statuses as *this* and *that* do.

In order to have a clear identification of noun phrases with abstract entities, I classify a NP as an abstract noun phrase if the head of the NP belongs to the following categories.<sup>5</sup>

- a. Verbal processing nominalization, usually acts of communication and having cognate illocutionary verbs, such as *accusation*, *admission*, *advice*, *announcement*, *suggestion*, *decision*, *warning*, etc.
- b. Language-activity nouns, referring to language activity but no illocutionary verbs, such as *ambiguity*, *comparison*, *consensus*, *contract*, *example*, etc.
- c. Nouns denoting abstract entities, such as *event*, *face*, *proposition*, etc.
- d. Mental process nouns, referring to cognitive states and processes and their results, such as *thoughts*, *belief*, *idea*, *hypothesis*, *finding*, *doubt*, *theory*, *thought*, *insight*, etc.
- e. Topic abstract nouns, referring to abstract nouns that cannot find referents that from previous text span, but referring to the abstract understanding of a whole text span, such as *topic*, *study*, etc.
- f. Inferable abstract nouns, referring to abstract nouns that have to be referred from word knowledge.

## 2.2 Forms of abstract referents

Asher (1993) claims that linguistic units that are possible abstract referents are from the following categories: gerund clause, *that*-clause, infinitive clause, verb phrase, nominal, noun-phrase that denote proposition-like entities, chunk of sentences. In this paper, I propose that NPs that denote abstract meaning are possible abstract referents as well, based on the following reasons.

First, NPs denoting abstract entities such as *a story* in (3a) can occur in the same predicate environment as other linguistic units used as possible

<sup>3</sup> Unless specified, examples in the paper are from ACE data.

<sup>4</sup> This example is from Asher (1993: 245).

<sup>5</sup> The categories in *a*, *b* and *d* are classification proposed by Francis (1994).

abstract referents do, such as the sentence *John left* in (3b).<sup>6</sup>

- (3) a. John left. I didn't believe it.  
 b. John told me a story. I didn't believe it.

Secondly, the frequent occurrences of the chain co-reference *this/that* and *it* gives further evidence, showing that *it* can refer to *this/that* which are noun phrases and denote abstract meaning. In (4), *this* refers to the event the reporter describes and *it* refers to the event as well via the co-reference with *this*.

- (4) Reporter: Hastert has risen quite high in republican ranks while keeping a fairly low profile.  
 Hastert : I didn't really seek **this** at all. **It** just kind of happened.

Thirdly, abstract noun phrases help anaphor resolution. The first abstract anaphor resolution system by Eckert and Strube (1999) model uses syntactic criteria to choose referents, and they point out that the system cannot deal with NPs denoting with abstract entities. Byron (2004) claims that her system can resolve abstract references to abstract NPs because her system is based on both syntactic and semantic criterion. However, the linguistic units that she uses as possible abstract referents do not include NPs denoting abstract entities.

Finally, in anaphora resolution, the classification of pronoun *it* is a very complicated case. Most anaphora resolution systems just manually picked up those cases in which *it* can be possible anaphors with NPs as their referents. Taking NPs denoting abstract entities as possible abstract referent helps to classify and resolve the pronoun *it* better by making use of the context information.

### 2.3 Distribution of abstract anaphors

Previous work on abstract anaphora has observed that *this* and *that* occur more frequently than *it* does as abstract anaphors. Webber (1991) finds that only 15.6% abstract anaphora in written English is using *it*, the larger percent is using *this* or *that*. Gundel et al. (2002) find the similar result.

<sup>6</sup> Example 3 is my own example.

Both of the above study only takes linguistic units larger than NPs as possible abstract anaphors. In this paper, the distribution is different from those claimed in the literature. I did a manual classification of all the occurrences of possible abstract anaphor forms from the data. Table 1 is the distribution of forms of anaphors found in the data<sup>7</sup>.

	that/that N	this/this N	It
Total	228	137	149
Structural	152	0	33
Individual	16	76	31
Abstract	52	49	49
Inferable	8	6	13
Unknown		6	23

Table 1 Distribution of forms of anaphors

*That/that N* and *it* are used more often as abstract anaphors than as individual anaphors, while *this/this N* are used more often as individual anaphors than as abstract anaphors.

The forms of abstract anaphors are the first criteria that can be used to classify abstract anaphors and individual anaphors. NPs denoting individual entities are NPs with concrete nouns as heads, such as *this house*, *this year*, *this person*, *this plane*, etc. NPs denoting abstract entities are NPs with abstract nouns as heads. NPs whose head nouns are belonging to the classification described in 2. 1 are taken as abstract NPs, such as *that event*, *this suggestion*, *this observation*, etc.

## 3 Classification of anaphors

Asher (1993) claims that the abstract referent types are encoded in the predicates of the abstract anaphors. For an anaphora resolution system, this is a crucial issue that needs to be taken care of in the resolving procedure. Based on Asher's assumption, both Eckert and Strube (1999) and Byron (2004) make use of predicate information of anaphors to classify individual anaphors and abstract anaphors.

### 3.1 Why classify anaphors?

First, anaphor forms can be either individual ana-

<sup>7</sup> Structural refers to the linguistic forms merely as structural purpose, such as that in that-clause. Unknown refers to the linguistic forms either denoting deictic meaning rather than anaphoric meaning or without clear antecedents.

phors or abstract anaphors, as in (5).<sup>8</sup>

- (5) a. Fitzgerald's translation of the Illiad is a masterpiece. It is over 300 pages long.  
b. Fitzgerald's translation of the Illiad is a masterpiece. It was a painful process.

The nominal *Fitzgerald's translation of the Illiad* refers to a collection of English sentences in (5a) and a process of translation in (5b). Predicates are helpful in disambiguating.

Secondly, it is possible to classify anaphors by their governing verbs. Cornish (1986) claims that the distinction between nominal *it*, *this*, and *that* and non-nominal *it*, *this* and *that* in English is signaled primarily by the semantic restriction imposed upon the pronoun by their governing verbs, as in following examples:<sup>9</sup>

- (6) a. It is impossible.  
b. It is heavy.  
(7) a. John believed it.  
b. John gave it to Harry.

Finally, a linguistic unit might be potentially several referent types. Predicate information can help to disambiguate referent types and classify anaphors as well. Consider (8).<sup>10</sup>

- (8) [John [crashed the car]<sub>j</sub>].  
a. This<sub>i</sub> annoyed his parents. (event)  
b. Jane did that<sub>j</sub>, too. (concept)  
c. This<sub>i</sub> shows how careless he is. (fact)  
d. His girlfriend couldn't believe it<sub>i</sub>. (proposition)

The sentence *John crashed the car* can refer to several ambiguous referent types, and the predicate of each sentence from *a* to *d* serves as a hint to specify which event type that the sentence refers to.

### 3.2 Asher's lists of predicates of abstract anaphors

Asher (1993) doesn't classify predicates for individual anaphors and abstract anaphors. However, he claims clearly that different referent types are

encoded in predicate information of abstract anaphors and he lists various sorts of verbs as the contexts for different referent types.

For propositional contexts, there are verbs for pure proposition, such as *true*, *false*, and attitude verbs, including positive factive (*believe*, *be certain*, *know*), nonfactive epistemic (*doubt*, *deny*). There are other verbs for projective proposition, such as rogatives (*wonder*, *ask*), buletics (*want*, *desire*), command verbs (*command*, *plead*), and permission verbs (*allow*, *permit*).

For factual contexts, there are verbs such as *indicate*, *show*, etc.

### 3.3 Eckert and Strube's predicate lists

Following Asher's assumption that the predicate of an abstract anaphor determines the referent type, Eckert and Strube (1999) encompass both individual entities and abstract entities in their anaphora resolution system by making use of the predication context of each pronoun and demonstrative to determine whether it is compatible with an individual entity or an abstract entity. They include some predicate type preferentially related to abstract anaphora in a list:

- Equating constructions where a pronominal referent is equated with an abstract object, e.g., *x is making it easy*, *x is a suggestion*.
- Copula constructions whose adjectives can only be applied to abstract entities, e.g., *x is true*, *x is false*, *x is correct*, *x is right*, *x isn't right*.
- Arguments of verbs describing propositional attitude which only take S'-complements, e.g., *assume*.
- Object of *do*.
- Predicate or anaphoric referent is a "reason", e.g., *x is because I like her*, *x is why he's late*.

They also include a list of predicate types that are incompatible to abstract entities but preferable to individual entities.

- Equating constructions where a pronominal referent is equated with a concrete individual referent, e.g., *x is a car*.

<sup>8</sup> Example (5) is from Asher (1993: 150).

<sup>9</sup> The two examples are from Cornish (1986: 70).

<sup>10</sup> Example 8 is from Eckert and Strube (2000: 58).

- Copula constructions whose adjectives can only be applied to concrete entities, e.g., *x is expensive*, *x is tasty*, *x is loud*.
- Arguments of verbs describing physical contact stimulation, which cannot be used metaphorically, e.g., *break x*, *smash x*, *eat x*, *drink x*, *smell x* but NOT *\*see x*

### 3.4 Byron's semantic constraints

Invoked by Eckert and Strube's system, Byron (2004) steps further. Rather than two simple preferable lists, Byron creates a set of most straightforward semantic constraints:

- 1) Verb semantic restrictions: some surface verbs are mapped to verb senses. Each verb is associated with a set of allowable semantic types for each of its argument position. For example, *looks/look like*, *sound*, *seems to be* are mapped to the verb sense: *Appears-to-be*. *Happen* is mapped to the verb sense *Happen*. For the verb sense *Happen*, its argument position is *Theme* and its semantic type restriction is *Event*. In the sentence *That happened last night*, the predicate verb *happen* restricts *that* to refer to an event.
- 2) Projective Proposition: the verb sense of infinitive complements of verbs such as *need/want/like* restricts the subject argument rather than *need/want/like*. For example, in *It needs to arrive by 1pm*, the referent of *it* is determined by *arrive* rather than by *it*.
- 3) Predicate complements: Complements of copula constrain the item in subject position. For example, in *That's true*. The predicate *true* allows only a proposition. So *that* refers to a proposition. In *That is where I grew up*. The *wh*-word *where* restricts the subject *that* to be a spatial location.

Compared to Eckert and Strube's two preference lists, Byron's system is more advanced. She develops the details on how the predication contexts constraint abstract anaphora resolution. However, Byron's verb sense only covers 73 verbs from spoken corpora. A lot of verbs that are useful in abstract anaphora resolution, such as *indicate*, *show*, are not included.

### 3.5 My preliminary finding and further plan

Given the importance of predicate information in abstract anaphora resolution, my main purpose is to extract the statistic relation between predicate context and anaphor classification from large corpora, which can be used in automatic abstract anaphora resolution.

So far, I labeled the data used in the paper, but this size is not big enough. After examining the small corpus, I found some predicate words are used for abstract anaphors only, such as *be involved in*, *be why*, *be right*, *be what*, *do*, *be good*, *make sense*, *happen*, *show*, *be understood*, *be done*, *be due to*, *be a task*, etc. There are also some predicate words used for individual anaphors only, such as *lower than*, *lose*, *receive*, *donate*, etc. There are some words have very strong preference. There are some predicate words have no preference and can be predicates of both type, such as *require*, *be helpful*, etc.

Since this is only a preliminary small sample corpus study, it is not worthwhile to do any statistical analysis on the relation between predicate information and anaphor types. I will use a large size of sample corpora to do the statistical study. Before that, following Byron (2004), a list of word senses will be created. Byron's word lists only cover 73 verbs denoting either individual entity context or abstract entity context from spoken corpora, and my list will cover a much larger number of verbs as well as nouns and adjectives as complements of copula. My goal is to come up with statistical information between predicate words and anaphor classification from large corpora, which is helpful in implementation of both individual anaphora resolution and abstract anaphora resolution.

## 4 Conclusion and future work

This paper intends to provide a corpus-study for abstract anaphora resolution and aims to function as a bridge between theoretical work and implementation in abstract anaphora resolution. To get a consistent analysis of abstract anaphors, I propose that NPs denoting abstract entities are possible abstract referents as well. I claim that the forms of anaphors, especially types of nouns in *this N* and *that N* are the first criteria in anaphor classifica-

tion. Following the previous study, I assume that the predicate information is the second criteria for abstract anaphor classification. I manually labeled the predicate words and found some words do have effect in classifying individual anaphors and abstract anaphors.

Further work will be done in the following aspects: first, use a much larger size of sample corpora and create a list of verb senses, and then do a statistical analysis to see how this list can be used in classification. Secondly, since the ultimate goal is to do anaphora resolution, I will further study two other aspects in abstract anaphora resolution. One aspect is to study the referent types and how different referent types and other clues in linguistic units used as abstract referent can be used in anaphora resolution. The other aspect is to study how discourse relations can help in abstract anaphor resolution as claimed in the literature (Webber, 1991; Asher, 1993).

## References

- Asher, Nicholas. 1993. *Reference to abstract Objects in Discourse*. Kluwer Academic Publishers, Dordrecht.
- Byron, Donna. 2004. *Resolving pronominal reference to abstract entities*. Ph.D dissertation. University of Rochester.
- Cornish, Francis. 1986. *Anaphoric Relations in English and French*. Croom Helm, London.
- Eckert, Miriam and Strube, Michael. 1999. Resolving Discourse Deictic Anaphora in Dialogues. In *EACL '99: Proc. of the 9th Conference of the European Chapter of the Association for Computational Linguistics*. Bergen, Norway, June 8-12, pp.37-44.
- Miriam Eckert and Michael Strube. 2000. Dialogue Acts, Synchronising Units and Anaphora Resolution, *Journal of Semantics* 17, pp.51-89.
- Francis, Gill. 1994. Labelling Discourse: An aspect of nominal-group lexical cohesion. In *Advances in Written Text Analysis*, M.Coutthard (ed.). 83-101. London: Routledge.
- Gundel, Jeanette K., Nancy Hedberg, and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 69(2): 274-307.
- Gundel, Jeanette K., Nancy Hedberg, and Ron Zacharski. 2002. Pronouns Without Explicit Antecedents: How Do We Know When a Pronoun is Referential? Presented at *DAARC-4 (the Fourth Discourse*

*Anaphora and Anaphor Resolution Colloquium*), Lisbon, Portugal.

- Mitkov, Ruslan. 2002. *Anaphora resolution*. Studies in Language and Linguistics. Longman/Pearson Education, London, New York, Toronto, Sydney, Tokyo, Singapore, Hong Kong, Cape Town, New Delhi, Madrid, Paris, Amsterdam, Munich, Milan, Stockholm.
- Webber, Bonnie L. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107-135.