Semantic Annotation of Anaphoric Links in Language

Kiyong Lee

Department of Linguistics, College of Liberal Arts, Korea University, Seoul, South Korea

Abstract The main purpose of this paper is to make some linguistic contributions to an ISO initiative to formulate a standard on the semantic annotation framework of reference, coreference, and other types of anaphoric phenomena in natural language. For this purpose, we first make a brief review of some of the existing coreference annotation (CA) schemes. We then formulate an abstract syntax ASyn\textsubscript{ana} for anaphoric annotation on which a variety of concrete syntaxes such as an XML-based concrete syntax CSyn\textsubscript{ana} can be developed to provide an interoperable representation format for the annotation. To satisfy the semantic adequacy of the proposed abstract syntax even partially, we check the possibility of developing formal semantics based on it. Such a semantics may be accepted as a valid application of the proposed annotation scheme. We finally consider the multilingual applicability of proposed ASyn\textsubscript{ana} by applying it to Korean, a non-inflectional agglutinating language with pro-drop properties.

Keywords Abstract Syntax, Anaphoric Link, Annotation, Concrete Syntax, Coreference, Formal Semantics, Semantic Annotation

1 Introduction

Datasets are important for any linguistic analysis at least in two respects.

1. Datasets make the analysis data-oriented with its focus on real data.

2. Datasets provide a basis of validating theoretical claims.\(^1\)

This paper thus presents two sorts of datasets of language fragments: one sort of datasets in English and another in Korean. The English datasets are used to carry the general discussion of anaphoric links in language, whereas the Korean datasets are used to check how it applies to non-European languages like Korean, a non-inflectional language known as an agglutinating language with pro-drop properties.

While the Korean datasets are introduced in Section 7 Application to Korean, here is a fragment of the English dataset as in (2), a somewhat bizarre story from an undisclosed, but freely available source:\(^2\)

(2) Not long ago {a retired professor\textsubscript{mt1} of a small university\textsubscript{mt2} in Essex\textsubscript{mt3}}\textsubscript{mt4} lost \{his\textsubscript{mt5} wealthy wife\textsubscript{mt6} who\textsubscript{mt7} had been ill for \{over ten years\}\textsubscript{mt8}. After \{her\textsubscript{mt9} death\textsubscript{mt10}} of the professor\textsubscript{mt11} and \{his\textsubscript{mt12} only daughter\textsubscript{mt13} mt14} found \{a large sum of money\}\textsubscript{mt15} left for them\textsubscript{mt16} to share.

Billy Jones\textsubscript{mt17}, 84, wanted to remarry and found \{a charming young Vietnamese woman\textsubscript{mt18} \textsubscript{mt19} aged 24, 40 years younger than his\textsubscript{mt20} own daughter\textsubscript{mt21}. He\textsubscript{mt22} told her\textsubscript{mt23} that, if she\textsubscript{mt24} agreed to marry him\textsubscript{mt25}, she\textsubscript{mt26} would receive \{eighty million pounds\}\textsubscript{mt27} \textsubscript{mt28}\{a million pounds\}\textsubscript{mt29} a year\textsubscript{mt30} right after their\textsubscript{mt31} marriage\textsubscript{mt32} and also inherit the rest\textsubscript{mt33} after his\textsubscript{mt34} death\textsubscript{mt35}. The lady\textsubscript{mt36} refused his\textsubscript{mt37} tempting offer\textsubscript{mt38}, instead marrying \{a young British bandmaster\textsubscript{mt39}}\textsubscript{mt40} with \{no promising career\textsubscript{mt41}}\textsubscript{mt42}.

This short story contains at least 43 noun phrases (NPs) of various forms that mention or refer to something, called discourse entity, in the universe of discourse, a non-empty set of discourse referents (see [38] and [39]). Each of these phrases, which are technically called mentions, is uniquely identified with an integer and its prefix mt that stands for mention in the text.\(^3\)

Most of the mentions are referentially grounded, namely referring expressions, and some of them are also identified as having coreferentially or, in more general terms, anaphorically related to others. To understand the whole story told by this short item, one should recognize these coreferential relations among the so-called mentions. The noun phrase A former head teacher\textsubscript{mt1} or the larger phrase a retire professor\textsubscript{mt1} and the name Billy Jones\textsubscript{mt19} for instance, refer to the same

\(^1\)This paper focuses on the design aspect of semantic annotation schemes based on the notion of abstract syntax validated with some consistent semantic interpretation of annotations. It does not present any corpus-based statistical analysis nor follows any big data approach to check the efficiency or other computational properties of an annotation scheme proposed.

\(^2\)Its grammaticality and relevance were checked by two native speakers of English, David Lovisek and Gary Clay Rector.

\(^3\)Not every mention refers to an (individual) entity (e.g., no man).
person and so do the noun phrase a charming young Vietnamese woman as and the noun phrase The lady.

Each of such pairs that refer to the same entity in the discourse referents consists of two mentions, one called ‘anaphor’ and the other its ‘antecedent’. They are also said to co-refer or be coreferential.

The antecedent of an anaphor may be split into more than one. The pronoun their in the first paragraph of the story, for instance, has its antecedent split into two mentions, a retired professor or the professor and his wealthy wife. There are half a dozen occurrences of pronouns such as his, her, he, she, their, and them that refer to either Billy Jones, his wife, their daughter or the Vietnamese lady. The two pronouns their and them have different sets of antecedents: the first one has professor and wife, their daughter and the Vietnamese lady. Only if necessary, this difference can be marked up as is being done by [30].

The study of reference, coreference, and other anaphoric relations has a long history with a very long list of literature. In the past several decades alone, a tremendous amount of work has been done in the area of syntax and semantics: generative grammar, for instance, has proposed various structural conditions and constraints on pronominalization and reflexivization, while formal semantics such as Montagovian semantics adopted model-theoretic denotational approaches, especially to the treatment of general quantification and variable binding. To cite a few, [77] started discussing constraints on variables that appear transformational rules in syntax and [49] introduced the structural notion of command between a referring expression and an anaphoric expression that played an important role in determining coreference and other related relations. [60] opened a new arena where both linguists and philosophers got engaged in analyzing and interpreting various types of quantifiers and pronouns or variables in language in the framework of higher-order formal logic and their scope and binding phenomena. This paper, however, has no intention to survey all this linguistic history that extends from syntax to semantics to pragmatics with discourse analysis, but focuses on a small area of most recent work by computational linguists on so-called coreference annotation (CA).

Reference or coreference resolution is a big issue in computational linguistics, especially in the area of information extraction (IE). There have been several important publications that deal with that issue: to cite some, we have: [30], [15], [23], [24], [72], [83], [84], [73], and [86]. We do refer to them at times, although coreference resolution (CR) differs technically from coreference annotation (CA) in which this paper focuses on.

While referring to these works and others to be cited, this paper aims at constructing a semantic annotation scheme (AS) for coreference and other anaphoric link phenomena in a language (English) that may be proposed as an ISO standard for language resources management.

The rest of the paper develops into seven sections. Section 2 reviews the five existing ASs, but in a non-critical way:

1. [5].
2. [85].
3. [30]’s MUC-7 Coreference Task Definition (CTD),
4. [64] and [65, 66]’s MATE/GNOME scheme revived by [82], and
5. [70]’s ISO-Space AS.

Section 3 shows how it adopts a two-level AS for anaphoric annotation, as followed by most of the anaphoric ASs. Given an input text, the first level simply identifies the whole set of referring or anaphoric expressions, called ‘mentions’, as possible markables. The second level is complex, consisting of three steps:

(3) a. Selecting possible anaphors from the set of possible markables.
   b. Paring each of the selected anaphors with their antecedent(s).
   c. Specifying the type of each anaphor-antecedent relation, called ‘anaphoric’ link.

Each of these three steps listed in (3a,b,c) is constrained and triggered by a small set of attribute-value specifications for each anaphor-antecedent pair of the mentions and each type of the anaphoric link.

Section 4 Formal description forms the main part of the paper that proposes (1) an integrated annotation scheme ASyn that for coreference and other anaphoric relations as an abstract syntax and (2) an XML-based concrete syntax CsynX derived from it with concrete attribute-value specifications. It is followed by section 5 Illustrations.

Section 6 shows how semantic forms are compositionally derived from each of the XML-represented annotations (XML-elements). This semantic section is an attempt to prove the semantic adequacy of the proposed ASyn at least in informal terms. This task is required by the general semantic principles on semantic annotation frameworks which are laid down in [36].

4 [70] calls this a case of split antecedent, whereas [78] views this as an instance of multiple antecedent.
Section 7 Application to Korean is added to the original paper in order to test the multilingual extensibility of ASyn\textsubscript{ana} with a possible modification of attribute-value specifications for a concrete syntax based on it. No corpus-based analysis is, however, presented, for that task alone requires a series of reports.

The abstract syntax of ASyn\textsubscript{ana} with its concrete syntax CSyn\textsubscript{ana}\textsubscript{max} presented in this paper is an extensively revised version of that presented by [54] with more formal rigor. Section 7 is a totally new addition to [54]. Section 8 concludes the paper with some summarizing remarks.

2 Review of Existing Annotation Schemes

2.1 Preliminaries

There are several types of the anaphoric relation. The best known type is coreference, an equivalence (symmetric, transitive, and reflexive) relation, that holds between two terms, called ‘mentions’, if and only if their denotations are identical.

(4) Definition-1: Coreference

Two terms $t_1$ and $t_2$ corefer iff $|t_1|^M = |t_2|^M$, where $|t_i|^M$ is the denotation or referent of a term $t_i$ with respect to a model $M$.

Coreference and binding are two different, but related linguistic phenomena, often discussed together. Consider examples (5a,b):

(5) a. John loves his mother. [coreference]
    b. Everyone loves his mother. [binding]

In (5a), the pronoun *his* may be understood as coreferring with the name *John* as its antecedent. In (5b), on the other hand, the pronoun *his* does not corefer with the quantified noun phrase *everyone*, but is treated in formal semantics as a variable bound by the universal quantifier. This paper attempts to accommodate both the type of coreference, as defined in (4), and some other types of anaphoric phenomena, such as the one illustrated in (5b), into the proposed ASyn\textsubscript{ana} with an XML-based concrete syntax CSyn\textsubscript{ana}\textsubscript{max}.

2.2 Coindexing

In linguistics, coreference and binding are both annotated in the same manner by coindexing, as shown below.

(6) a. John, loves his mother. [coreference]
    b. Everyone, loves his, mother. [binding]

Split antecedents can also be represented by coindexing with a set index such as \{x,y\}. Here are some examples, (7a,b,c), taken from [78]:

(7) a. Mary\textsubscript{1} told John\textsubscript{2} that they\textsubscript{1,2} should invest in the stock market.
    b. Every woman\textsubscript{1} told [her\textsubscript{1} husband\textsubscript{2}] that they\textsubscript{1,2} should invest in the stock market.
    c. Every man\textsubscript{1} told [each of his\textsubscript{1} girlfriends\textsubscript{2}] that they\textsubscript{1,2} were going to get married.

Coindexing is not, however, expressively powerful enough to mark up details of anaphoric relations. The treatment of reciprocal pronouns is one of such cases.

(8) a. They\textsubscript{1} love each other.
    b. If everyone\textsubscript{1}, were to love one another\textsubscript{2}, then they\textsubscript{1} wouldn’t want me to make a sacrifice, ...

More than mere coindexing is called for an adequate interpretation of the anaphoric phenomena shown in (8).

2.3 The TEI-based Annotation Schemes

2.3.1 Overview

There are two almost identical ASs for coreference: [5] and [85]. They are both based on XML and also on the TEI Guidelines. They differ from each other mainly because [5] followed a much earlier version of the TEI Guidelines, while [85] is the most recent version, updated December 2016.

2.3.2 [5]

The AS, proposed by [5], consists of two (XML) elements: \(<rs/>\) for referring strings and \(<link/>\) for coreference. For each of the two elements, we can specify their associated attributes and possible values:

(9) a. Attributes for the Element \(<rs/>\):
    attributes = type, key;
    type = "object";
    key = ID;

b. Attributes for the Element \(<link/>\):
    attributes = type, targets;
    type = "coref";
    targets = IDREF IDREF.

Here is an example:

(10) \(<rs>Kill <rs type="object" key="01">an active, plump chicken</rs/>. Prepare <rs type="object" key="02">it</rs> for the oven, cut <rs type="object" key="03">it</rs> into <rs type="object" key="04">four pieces</rs/> and roast <rs type="object" key="05">it</rs> with thyme for 1 hour.</rs/>\)

\(<link type="coref" targets="02 01"/>\)
\(<link type="coref" targets="03 02"/>\)
\(<link type="coref" targets="04 03"/>\)
\(<link type="coref" targets="05 04"/>\)

\(^6^\)As a specification language, we adopt [31] Extended BNF (Backus-Naur Form) that can be converted into a document type definition (DTD).

\(^7^\) [5] annotated a French version of the original English example (16) that occurs in [4], page 202.
The chain of coreference relations may not preserve the original identity. In example (10), the live chicken was killed and cut into four pieces and became roasted chicken.

2.3.3 [85]

[85] (16.5.1) discusses correspondence between textual segments:

(11) <title xml:id="SHIRLEY">Shirley</title>, which made its Friday night debut only a month ago, was not listed on <name xml:id="NBC">NBC</name>’s new schedule, although <seg corresp="#NBC" xml:id="NETWORK">the network</seg> says <seg corresp="#SHIRLEY" xml:id="SHOW">the show still is being considered.</seg>

In this textual fragment, the name Shirley is annotated as the title of a show being broadcast over NBC, a television network. The text contains no pronominal forms, but the two nominal forms, the show and the network, are understood as corresponding to the two names Shirley and NBC, respectively. The annotation of the two segments, namely those nominal forms, right above introduces the attribute @corresp to indicate such a coreferential relation for each of the two.

As shown by (12), the use of the elements <linkGrp> and <link> makes correspondence relations more explicit:

(12) <linkGrp type="anaphoric_link" targFunc="antecedent anaphor"> <link target="#Shirley #show"/> <link target="#NBC #network"/> </linkGrp>

The annotation (12) represents two instances of the anaphoric link involving anaphors and their antecedents. The attribute @target has two arguments, as specified by the attribute @targFunc: the first argument is antecedent and the second anaphor. The element <linkGrp> allows several instances of the anaphoric link to be grouped together, while simplifying the specification of the element <link> with a single attribute @target.

2.4 The MUC-7 Coreference Task Definition

2.4.1 Overview

[30]’s CTD (Coreference Task Definition) lists four purposes of constructing an AS in the order of their importance. The first two are: (1) to support the MUC (Message Understanding Conference) information extraction tasks and (2) to be able to achieve good (ca. 90%) inter-annotator agreement.

Creation of a corpus for research on coreference and discourse phenomena is the last goal.

The MUC projects were taken up by [17] of NIST (National Institute of Standards and Technology) and other organizations working on linguistic data. [57] then produced detailed annotation guidelines on the annotation of entities and locations with the specification of their types. [1] then recently compared the annotation works, ACE, ERE, TAC-KBP, and FrameNet annotation standards focusing on event annotation. Besides the annotation of entities, the research interest in event annotation was kept by works such as [2]. The report by [67] introduces the CoNLL2011 project with OntoNotes. In this paper, we just focus on [30] to see how anaphoric annotation started in earlier stages.

2.4.2 Markables

[30]’s CTD restricts the set of its markables to nouns, that is, names, noun phrases, or pronouns. Noun phrases include dates (January 23), currency expressions ($1.2 billion), percentages (17 %), and temperatures (70 degrees) that contain numerical values. Possessive, demonstrative, and reflexive pronouns are markables. So are the first person pronouns. The interrogative pronouns (who, which engine) are not markables.

In [30], verbs and other verbal forms such as gerunds (Slowing the economy) are not markables. Implicit pronouns, that is, null anaphora (Bill, called John and e, spoke with him for an hour.) and presumptive or intrusive pronouns (the movie, which I saw t) as well as relative pronouns (complementizers) are not treated as markables.

2.4.3 Extents

The extent of a markable is a maximal string, while its head is marked with an attribute MIN (minimal string). The maximal noun phrases thus include their modifiers, appositional phrases, non-restrictive relative clauses, and prepositional phrases (Fred Frosty, the ice cream king of Tyson’s Corner, MIN="Fred Frosty").

2.4.4 Coreference Links

Coreference in [30] is not restricted to referential identity. Here is the general principle for the annotation of coreference that they proposed:

(13) Definition 2: Coreferential

Two markables are coreferential if they both refer to sets, and the sets are identical, or they both refer to types, and the types are identical.

The principle (13) thus allows the possible coreferentiality between bound anaphora and quantified NPs that are their antecedents.

8The original annotation of the French version given in [5] was modified to suit the English example (10).
9The hash sign # indicates that each item preceded by # such as #NBC and #Shirley in example (11) has occurred somewhere in the annotated fragment. The #NBC which is given as a value of the attribute @corresp in the element <seg> occurs in the <name xml:id="NBC">.

10All of the examples given in subsection 2.4 are copied from [30].
11The phrases program trading, excessive trading, slowing of the economy are noun-like, so they are treated as markables.
12See [16], [19], [20], and [88] for detailed discussions of interpreting various uses of pronouns and [80] for presumptive and intrusive pronouns.
Examples (14) are given by [30] to illustrate bound anaphoric relations.13

(14) a. {Most computational linguists}; prefer their; own parsers.
b. {Every TV network}; reported its; profits yesterday. They; plan to release full quarterly statements tomorrow.

2.4.5 An SGML Serialization

[30] represents its coreference annotation in SGML. It introduces only one element <COREF> for the annotation of markables and also of their coreferential link type IDENT with the following specification of attribute-values:

(15) List of Attributes and Possible Values for <COREF>
ID=INTEGER;
MIN=CDATA; { * Head of the whole extent * }
REF=IDREF; { * Antecedent * }
TYPE=IDENT;
STATUS=OPT; { *if the reader is uncertain about the identity relation.* }

2.4.6 Illustrations

Examples (16) and (17) are copied from [30]:

(16) <COREF ID="100">Lawson Mardon Group Ltd.</COREF> said <COREF ID="101" TYPE=IDENT REF="100">it</COREF> ....

(17) Our <COREF ID="102" MIN="Board of Education">Board of Education</COREF> budget is just too high. The Mayor said <COREF ID="102" STATUS="OPT" TYPE="IDENT" REF="102">Livingstone Street</COREF> has lost control.

In example (17), the annotator is uncertain about the IDENT relation of the Board of Education and Livingstone Street, although they are locally identical. That is why STATUS="OPT" is introduced into the annotation (17). Conjoined noun phrases are treated as one extent in example (18).

(18) <COREF ID="1" MIN="boys and girls">The sleepy boys and girls</COREF> enjoy <COREF ID="2" REF="1" TYPE="IDENT">their</COREF> breakfast.

In example (19) there are two occurrences of president, but they are not annotated as being coreferential. <COREF ID="7">has neither REF nor TYPE specified as being IDENT.

(19) <COREF ID="5">Fred</COREF> resigned as <COREF ID="6" MIN="president" REF="5">president of IBM</COREF>; next month, <COREF ID="7">the president</COREF> will be <COREF ID="8" REF="7">Mary</COREF>

Bound anaphors are treated in (20).

(20) <COREF ID="1" MIN="man">every man who knows</COREF ID="2" REF="1" TYPE="IDENT">his</COREF> own mind</COREF>

The entire string every man who knows his own mind and the pronoun his are annotated as coreferring with their relation type being IDENT. There are many other illustrations, for instance, for the annotation of apposition, predicative nominals and time-dependent identity, types and tokens, functions and values, and metonymy.

2.5 The MMAX2 Multi-level Annotation Scheme

2.6 Overview

[79] use [61]'s GUI-based MMAX2 annotation tool for coreference resolution to build a fully coreference-annotated large corpus of 266 scholarly papers from the ACL anthology. Here we briefly introduce the MMAX2 coreference AS.

2.6.1 Markables

Only proper names, noun phrases, and pronouns are markables, called possible entity ‘mentions’. There are 8 mention types: (1) def-np (definite NPs), (2) pper (personal pronouns), (3) ne (proper names including citations), (4) ppos (possessive pronouns/determiners), (5) indef-np (indefinite NPs), (6) conj-np (coordinations), (7) pds (demonstrative pronouns), and (8) preflexive (reflexive pronouns).

Unlike [30], this AS treats relative pronouns (who, which, whose, that, ...) as markables and excludes bound anaphora (Every teacher), likes his, job) as well as predicative nominals ([A mason], is a workman).13

2.6.2 Anaphoric Links

Following [86] and their two other related works, [47, 48], this AS differentiates coreference from other types of the anaphoric link. It suggests that the annotation of coreference proper be separated from other tasks such as annotation of bound anaphors and of the relation between a subject and a predicative NP. It calls for a division of labor that achieves better inter-annotator agreement.

2.7 The MATE Annotation Scheme

2.7.1 Overview

The MATE annotation scheme for anaphoric relations has been developed mainly by [64], [65], and [66]. It has accommodated several annotation schemes including the two
annotation schemes, the TEI-based AS by [5] discussed in 2.3 and the MUC-7 CDT by [30] discussed in 2.4. Its application domain extends from ordinary text to dialogues to discourses with the GNOME corpus. Its semantic theory, especially related to discourse modeling, is solidly based on [87] and [39]. The MATE website has, however, been shut down, whereas [82] attempted to restore it by introducing a short summary of the MATE annotation scheme.

2.7.2 Example

Here is an example, taken from [64] and also from [65], illustrating how the MATE annotation scheme works.

(21) a. 5.12 : we're gonna take the engine E3
15.13 : and shove it over to Coming
15.14 : hook it up to the tanker car
15.15 : and send it back to Elmira

b. coref, xml :

15.12 : we're gonna take
<de ID="de_15">the engine E3</de>
15.13 : and shove <de ID="de_16"> it</de> over to Corning
15.14 : hook <de ID="de_17"> it</de> up to
<de ID="de_18"> the tanker car</de>
15.15 : and
15.16 : and send <de ID="de_19"> it</de> back to Elmira

The XML element, named <de>, stands for discourse entities that referred to by referring and anaphoric expressions. These expressions are markables. In example (21), the links, <link>, that capture anaphoric relations are represented standoff as in (b), while the markables are inline annotated as in (a).

2.7.3 Attribute-Value Specifications

As illustrated with 2.7.2, the MATE annotation scheme in XML has two elements: <de> and <link>. The element <de> annotates expressions that refer to discourse entities, specifying its attribute @ID with a unique value referring to each of the markables such as noun phrases and pronouns. The element <link> specifies anaphoric relations, containing one or more <anchor> elements.

Both <link> and <anchor> have an attribute @href. The value of the attribute @href for <link> refers to an anaphoric expression, whereas the value of @href for <anchor> refers to an antecedent. As shown in the last <link>, an anaphoric expression may have more than one antecedent or an ambiguous antecedent.14

The element <link> has an attribute @type. The list of possible values includes ident for coreference and subset denoting a subset of the set denoted by a discourse entity. This list can be extended to include values like part by following the DRAMA scheme by [63] as mentioned in [66], Extended Relations.

2.8 The Brandeis ISO-Space Annotation Guidelines

The ISO-Space Working Group at Brandeis University produced a manual of annotation guidelines for spatial information [70] and proposed it as an annex of an earlier CD (committee draft) version of [35], an ISO international standard for spatial annotation. First, the annex of a CD version annotates spatial entities (I am sitting in the car se).15 as referring expressions. Second, it introduces an element <metaLink> to annotate the three different types of coreference between these spatial entities:

(22) a. coreference,
b. subcoreference, and
c. split coreference.

Third, it specifies a set of attributes and their possible values for the element <metaLink>, as shown in (23):

(23) Attributes and Possible Values for <metaLink>

attributes = id, [objectId1], [objectId2], relType, [comment];
{* the attributes in square brackets are implied.*}

id = "meta"INTEGER;
ObjectId1 = IDREF;
ObjectId2 = IDREF;
relType = "coref", "subCoref", "splitCoref";
comment = CDATA;

14 As specified in [33] or [85], the element <vAlt> for alternate values can represent ambiguous cases in a more explicit or correct way.
15 A spatial entity is introduced into [35] as an element of a type of entity that participates in location-involving motions or non-motion events. Its identifier is marked with a prefix se.
Instead of using XML as its representation language, [70] adopts the predicate-logic-like forms to represent the three different types of coreference. Examples are given in (24):

(24)  a. \{Two cars\}_{se1} are on the street. One\_{se2} of them\_{se3} turns left.

b. spatialEntity(id=se1, extent=two cars, countable=yes)
   spatialEntity(id=se2, extent=one, countable=yes)
   spatialEntity(id=se3, extent=them, countable=yes)
   metaLink(id=meta1, objetID1=se1, objectID2=se3, relType=coref)
   metaLink(id=meta2, objetID1=se1, objectID2=se2, relType=subCoref)

The list of annotations in (24) can be interpreted as stating that the referents of two spatial entities cars\_{se1} and them\_{se3} are identical, while one\_{se2} partially coresfers with the spatial entity type expressions cars\_{se1}.

Here is example (25):

(25)  a. \{John\}_{se6} and Mary\_{se7} went shopping.

b. metaLink(id=meta4 objectID1=se8 objectID2=se9 relType=coref)
   metaLink(id=meta3 objectID1=se7 objectID2=se8 relType=subCoref)
   metaLink(id=meta4 objectID1=se9 objectID2=se6 relType=splitCoref)
   metaLink(id=meta5 objectID1=se9 objectID2=se7 relType=splitCoref)

The names John\_{se6} and Mary\_{se7} in (25) are each treated as a referring expression. At the same time, the whole phrase \{John\_{se6} and Mary\_{se7}\}_{se8} as a group is also treated as a referring expression. And then the antecedent of the plural pronoun They\_{se9} is split into two: John\_{se6} and Mary\_{se7}.

3 Steps of Annotating Coreference and Other Anaphoric Relations

3.1 Overview

Given an input text, the task of coreference or other anaphoric link annotation is three-fold:

(26)  Step A: Identification of a set of mentions in the text that refer to something in the domain of discourse referents as its markables,
   Step B-1: Identification of a set of anaphor-antecedent pairs of the mentions that are anaphorically related and
   Step B-2: Specification of the type of such a relation.

To trigger and constrain these annotation steps, the entity type of mentions and the anaphoric relation are assigned a set of required or implied attribute-value specifications. Step B-1 and Step B-2 constitute two sub-processes unified into one, for they depend on each other. The input text can be of any size. It can range from a short sentence to a very large corpus.

3.2 Identifying Markables and Extents

The set of possible markables consists of terms or mentions, which comprise both referring and non-referring expressions in a text. As attested quantitatively by various reference resolution experiments such as [15], [24], [83], and [71], these mentions are mostly noun phrases (NPs). Noun phrases are then categorized into four forms as shown in (27):

(27)  a. Proper names: e.g., Mary, Obama, The United States of America
    b. Definite or indefinite nominals with plurality and other agreement specifications: e.g., The boy, the girls, apples, oxen
    c. (Generalized) universal or existential quantifiers: e.g., every man, three women, some countries
    d. Definite or indefinite pronouns with gender and number specifications.

Definite and indefinite pronouns listed in (27d) are divided into four sub-classes as shown (28):

(28)  a. Personal pronouns: e.g., he, she, they, it,
    b. Reflexives: e.g., himself, herself, themselves,
    c. Reciprocals: e.g., each other, one another, and
    d. Demonstratives: e.g., this, these, those.\(^{16}\)

The list of features such as ‘gender’ and ‘number’ specifies what morphosyntactic features are required or implied for the identification of mentions. The annotation of these features could be done at earlier stages of annotating raw data such as tokenization and morphosyntactic annotation. The process, whether manual or automatic, of marking up these mentions as markables should be straightforward at this basic level.

3.3 Anaphoric Links

The main task of annotating coreference and other types of the anaphoric link is to recognize antecedent-anaphor pairs among the set of markables and also to identify the type of their anaphoric link.

3.3.1 Anaphor-antecedent Pairs

Anaphors are part of the set of mentions, being mostly pronouns and other pronominal forms (see [42]). They are thus easily identified.

(29)  a. Bob loves Jane, but she doesn’t love him.

\(^{16}\)Interrogatives are excluded.
b. Bob was tired, and so was I.

Some definite noun phrases can be anaphors, too. Here are some examples:

(30) a. {The project leader}$_i$ is refusing to help. {The jerk}$_i$ thinks only of himself.
   b. {Hilary Clinton}$_i$, {Bill’s wife}$_i$.

Among the list of pronouns, we may also include the use of *it* referring to propositions, facts, actions, etc., or the use of *so* that may involve so-called sloppy identities, as shown below:

(31) a. John said {he has been to heaven}$_i$, but I don’t believe {it}$_i$.
   b. John {loves his wife}, and so {does}$_i$2 Bob.

Examples such as these are often discussed in linguistic literature, but have been seldom treated in computational work.

The so-called expletive *it* and *there*, the complementizer *that*, and the impersonal use of the pronoun *it*, as shown in (32), are excluded from the list of possible anaphors as well as from the list of possible markables.

(32) a. *It*$_{exp}$’s impossible to go out now, for *it*$_{imp}$’s raining cats and dogs.
   b. *It*$_{exp}$ is also reported that$_{comp}$ *there*$_{exp}$ is a storm approaching from the south.

The identification of anaphors as well as mentions can also be triggered by the morphosyntactic features of markables.

3.3.2 Types of the Anaphoric Link

Unlike anaphors, antecedents can be of any class of a word, phrase, or clause. It should, however, be a subset of markables as specified by Step A of annotation, listed in (26). If verbal forms are excluded from the set of markables for some practical reasons, then they would not be in the set of possible anaphors or antecedents.

The extent of antecedents is not restricted to a single word or phrase, but may extend to larger phrases such as conjoined phrases:

(33) {The boys$_j$ and the girls$_j$}$_k$ met at a party and they$_k$ danced all night.

Antecedents may not be contiguous, either, but split into two or more phrases, as in:

(34) I met {a farmer}$_j$ and {his$_j$ dog}$_k$ and we$_{(i,j,k)}$ all walked together.

There are at least two uses of pronouns:

(35) a. anaphoric: *John$_i$* loves his$_i$ wife.
   b. indexical or deictic: Look at him$_1$. *He$_1$* is naked.

Context: the speaker pointing to a person over there.

In the anaphoric use, the pronoun *his$_i$* finds its antecedent *John$_i$* in the given text. In the indexical use, the antecedent of the pronoun *him$_1$* is not found in the text, but provided contextually.

Pronouns can be antecedents as well as anaphors in the chain of an anaphoric link.

(36) John$_i$ loves {his$_i$$_2$ wife}$_j$ and she$_j$ also loves him$_{i_1}$.

Anaphoric links may be forward or backward. The term that corefers with a pronoun normally precedes it, thus being called ‘antecedent’. This so-called antecedent may also come after its related anaphor, as in:

(37) When she$_j$ returned home, *Sue$_i$* was surprised to find her dog gone.

In such a case, the pronoun is often called ‘cataphor’.

Sometimes it is difficult to decide which is an anaphor and which is its antecedent, as especially in appositive cases (*Seoul*, {the capital of South Korea}$_j$, where *i* and *j* corefer). In such cases, we simply have to state that they correspond to or corefer with each other.

The antecedent-anaphor relation is normally a one-to-one relation, but there are cases in which the antecedent of an anaphor is split into many. Besides this case of split coreference, [70] lists subCoreference as another type of coreference:

(38) I have {two cars}$_i$, but one$_j$ of them$_i$ broke down.

Here one$_j$ is a member of the set of two cars$_i$.

4 Formal Description

4.1 Overview

[7] provides a formal description of the annotation structure, consisting of two levels of syntax:

(39) a. One is an abstract level of an annotation, called ‘abstract syntax’;
   b. Another level, a concrete level of representing annotations, called ‘concrete syntax’.

Every abstract syntax for semantic annotations must be supported by an explicit (formal) semantics, as sketched in section 6. An XML-serialization of an abstract syntax is an instance of a concrete syntax. The semantics of a concrete syntax is defined as the semantics of the abstract syntax for which it defines a concrete representation. (Different representations of the same abstract syntax thus have the same semantics.)

Following [7], we first sketch a general view of ‘abstract syntax’ in 4.2. We then follow a modified version of [55] to present a general abstract syntax ASyn in 4.3 and an abstract syntax for anaphoric links ASyn$_{ana}$ as a particular instantiation of the general abstract syntax ASyn in 4.4.
4.2 A Sketch of Abstract Syntax

According to [7] and two of his subsequent works, [8] and [10], the abstract syntax of an annotation scheme consists of two parts:

(40) a. A conceptual inventory that specifies the basic concepts from which annotation structures are built up;

b. A specification of the possible ways of combining elements of the conceptual inventory into annotation structures.

An annotation structure is a set, consisting of two kinds of elements: entity structures and link structures. Entity structures provide linguistic information about a region of primary data; link structures provide information about the semantic relation between regions of primary data.

In the case of annotating coreference and other anaphoric links, entity structures correspond to the entities that are related by anaphoric links, and link structures to the linkings of anaphoric expressions to their antecedents.

An entity structure is a pair \((m, a)\) where \(m\) is a markable that identifies a region of primary data, and \(a\) is the specification of the semantic information that the annotation provides about that region of primary data. In the abstract syntax \(\text{ASyn}_{\text{ana}}\) for anaphoric annotation, the \(a\) component of an entity structure is an \(n\)-tuple, \(3 \leq n \leq 7\), consisting maximally of a semantic type \(t\), a definiteness \(d\), a morphosyntactic form \(f\), a natural gender \(g\), a plurality \(p\), and a collectiveness \(c\) (more about these elements below). The fact that the length \(n\) of these \(n\)-tuples may vary reflects the optionality of some of the elements.

A link structure is a triplet \((\varepsilon, \eta, \rho)\), where

(41) a. \(\varepsilon\) is an entity structure and \(\eta\), a set of entity structures, corresponding to an ‘anaphor’ and a set of its ‘antecedents’, respectively;

b. \(\rho\) is a relation between a pair of entity structures or between an entity structure and a set of entity structures.

For the abstract syntax \(\text{ASyn}_{\text{ana}}\) for anaphoric annotation, the conceptual inventory is a 9-tuple \(\langle M, T, D, F, G, P, C, Q, R \rangle\), where

(42) a. \(M\) is a non-empty set of markables;

b. \(T\) is a set of semantic types;

c. \(D\) is a set of definiteness values;

d. \(F\) is a set of morphosyntactic forms;

e. \(G\) is a set of natural genders;

f. \(P\) is a set of singular/plural values;

g. \(C\) is a set of ‘collectivity values’;

h. \(Q\) is a set of generalized quantifiers and

i. \(R\) is a set of binary relations over the set of entity structures, corresponding to the various types of anaphoric links.

The annotation structures are defined by an assignment \(\@\) that specifies the semantic components of entity structures. For each markable \(m\) in \(M\), \(\@\(m\)\), an \(n\)-tuple, \(3 \leq n \leq 7\), of elements from \(T \times D \times F \times G \times P \times C \times Q \times R\).

We do, however, follow [52, 53], and his more recent work [55] in formalizing an abstract syntax in algebraic terms that are often used in defining formal grammars. We also use [37] extended BNF as a meta-language to specify various features of data types because BNF is expressively more powerful than simple set-theoretic listing.

4.3 General Abstract Syntax

Given a fragment \(L_i\) of a language as primary data for annotation, the general structure of an abstract syntax \(\text{ASyn}\) for an annotation structure can be formally defined as a triple:

(43) \(<M, T, \@\>\), where

1. \(M\) is a set of (possibly null or non-contiguous) strings of character segments, called ‘markables’, in \(L_i\), delimited by \(E_i\);

2. \(T\) is a finite set of semantic types, consisting of \(E_i\) and \(L_i\), where \(E_i\) is a finite set of ‘entity types’ and \(L_i\) is a finite set of the types of link over \(E_i\);

3. \(\@\) is a set of assignment functions from \(T\) to linguistic features.

For semantic annotation, its markables in \(M\) are strings of character segments which are identified as tokens, words or phrases in a fragment of a language, given as its primary data. This is so because semantic annotation normally presupposes that its input data has been preprocessed by word segmentation or morpho-syntactic analysis.

Empty strings are allowed as markables in \(M\). They represent so-called non-consuming tags with their use licensed in [35]. Consider example (44),

(44) a. John got up late and \(\emptyset_i\) had to skip breakfast.

b. We camped \(\emptyset_{pl1}\) near the river.

The missing subject in the right conjunct of example (44a) can be represented as a non-consuming empty tag, if necessary. The \(\emptyset_{pl1}\) in (44b) refers to some place near the river and the event of camping is anchored to that place \(pl_{1}\).

\footnote{The specification of morphosyntactic forms, and several of the other elements may refer to other levels of annotation.}

\footnote{Here we are assuming that the media type of markables is textual. The media type of markables can, however, be extended to other types than the textual type, such as visual objects (pictures and videos).}

\footnote{See Definition 3.11 “Terms and definitions” and Annex A.3.4 “Special section: Non-consuming tags”, pp.29-30.}
$T$ is a very small set, consisting of a very small number of entity types $E_t$ and link types $L_t$. As will be shown in 4.4, the sets $M, E_t,$ and $L_t$ are interrelated by the set $\@$ of feature assignment functions associated with those sets.\footnote{The term ‘feature’ is used here loosely and interchangeably with the term ‘property’.

A particular feature $\@_i$ in $\@$ assigns an entity type $\tau$ in $E_t$ and other associated features $F_i$ to each markable $m$ in $M$, generating an annotation structure $\langle m, \@_i \rangle$ such that $@_i(m) = t \oplus F_i$.

For each link type in $L_t$, $\@$ also defines a set of (binary) relations $\rho$ over the set $\langle M, E_t \rangle$ of entity structures. Each of the specifications can be represented in extended BNF [37] as a function $\@_i$ over $T$ that assigns a value type such as CDATA (character data) or IDREF (identifier reference) associated with XML to each of the features associated with each markable $m$ in $M$. The particular names of features and their value types mentioned in the specifications are not fixed, but may vary for each concrete syntax or dialect.

4.4 Abstract Syntax for Anaphoric Annotation

4.4.1 General

The abstract syntax $\text{ASyn}_{\text{ana}}$ for anaphoric annotation is a particular instantiation of the general abstract syntax $\text{ASyn}$ which is introduced in (4.3) General Abstract Syntax, $\langle M, T, \@ \rangle$.

(45) $\text{ASyn}_{\text{ana}}$ consists of:

1. $M$ is a set of (possibly null or non-contiguous) sequences of tokens or words that refer to objects of the entity types specified in $E_t$;
2. $T$ is a set of semantic types, consisting of two types: a single basic entity type and a single anaphoric link such that each link type is associated with a triple $\langle e, \eta, \rho \rangle$; and
3. $\@$ is a set of feature assignments to be specified for each of the types in $T$ in extended BNF separately.

Among the elements of the 9-tuple $(M, T, D, F, G, P, C, Q, R)$ introduced in (42), the first two elements $M$ and $T$ exactly correspond to the $M$ and $T$ in $\text{ASyn}_{\text{ana}}$. The other six elements, $D, F, G, P, C, Q,$ and $R$, are treated by the feature assignment function $\@$ which is associated with each of the types in $T$.

4.5 Feature Assignments $\@$

Associated with $\text{ASyn}_{\text{ana}}$, the feature assignment $\@_{\text{entity}}$ assigns to each entity, which is referred to by a markable, a value for each of the attributes specified by $D, F, G, P, C, Q,$ and $R$, as shown (46):

(46) Feature Assignment $\@_{\text{entity}}$:

\[
\text{attributes} = \text{identifier}, \text{target}, \text{type}, \text{form}, \text{[definite]}, \text{[natural Gender]}, \text{[plural]}, \text{[collective]}, \text{[quantified]}, \text{[comment]};
\]

\{\text{The attributes in square brackets are optional attributes.}\}

\text{identifier} = \text{unique ID};
\text{target} = \text{IDREF|NULL};
\text{definite} = \text{BOOLEAN};
\text{naturalGender} = \text{CDATA};
\text{plural} = \text{BOOLEAN};
\text{collective} = \text{BOOLEAN};
\text{quantified} = \text{CDATA};
\text{comment} = \text{CDATA};

The feature assignment function associated with the type link is specified as in (47):

(47) Feature Assignment $\@_{\text{link}}$:

\[
\text{attributes} = \text{identifier}, \text{anaphor}, \text{antecedents}, \text{relType};
\text{identifier} = \text{unique ID};
\text{anaphor} = \text{IDREF};
\text{antecedents} = \text{IDREFS};
\text{relType} = \text{CDATA};
\]

The feature assignments (46) and (47) specify only the data type of a value for each attribute, which is either required or implied, without providing any specific values. Specific values are given by concrete syntaxes based on the abstract syntax for anaphoric annotation.

4.6 XML-based Concrete Syntax

We introduce an XML-based concrete syntax $\text{CSyn}_{\text{anaX}}$, corresponding to the abstract syntax $\text{ASyn}_{\text{ana}}$ for anaphoric annotation. First, the annotation of each of the entity or link structures is represented as an XML element. $\text{CSyn}_{\text{anaX}}$ introduces two elements $\langle \text{entity} \rangle$ and $\langle \text{anaLink} \rangle$ that correspond to each of the entity structures and also to each of the link structures, respectively, as defined in $\text{ASyn}_{\text{ana}}$. Both of these XML elements have an $\@_{\text{identifier}}$ attribute in order to allow references from within the representation of a certain link structure to the representations of specific entity structures or other link structures. Moreover, $\langle \text{entity} \rangle$ structures have a $\@_{\text{target}}$ attribute for representing the markables that they associate linguistic information with.

Second, the assignment $\@$ for $M$ can be transduced into $\text{CSyn}_{\text{anaX}}$ as specified in (48):

(48) Attributes and Values for the element $\langle \text{entity} \rangle$:

\[
\text{attributes} = \text{identifier}, \text{target}, \text{type, form, [definite], [natural Gender]};
\]

\text{The element $\langle \text{anaLink} \rangle$ may be understood as standing for $\langle \text{link type="anaphoric"} \rangle$.}
The attribute name @ground may have more than one antecedent as its value. This value can be a (possibly null or non-contiguous) sequence of tokens or their IDs.

By allowing the value quant as a possible value for the attribute form, CSyn\textsubscript{anaX} treats the anaphoric link between a quantifier and a bound anaphor (e.g., pronoun). As defined in (48), CSyn\textsubscript{anaX} may treat reciprocals (e.g., each other, one another).

The feature assignment function for \textless anaLink\textgreater is specified as in (49):

\begin{verbatim}
(49) Attributes and Values for the element <anaLink>
attributes = identifier, figure, ground, relType, [comment];
identifier = IDPrefix="x" followed by an INTEGER;\textsuperscript{23}
target = IDREF|CDATA;\textsuperscript{24}
type = "person"|"organization"|CDATA;\textsuperscript{25}
form = "name"|"nom"|"pron"|"quant";\textsuperscript{26}
definite = "yes";\textsuperscript{27}
naturalGender = "male"|"female"|CDATA;\textsuperscript{28}
plural = "yes";
collective = "yes";
quant = "universal"|CDATA;\textsuperscript{29}
comment = CDATA;
\end{verbatim}

The attribute name @figure stands for an anaphor, while the attribute name @ground stands for the entity type antecedent.\textsuperscript{31} The attribute @relType introduces values other than ident for referential identity. These values allow

\begin{itemize}
\item \texttt{ident}: referential identity either at an individual level or at a set or group level;
\item \texttt{partIdent}: referential partial identity or subcoreference;
\item \texttt{splitIdent}: antecedents are split into more than one;
\item \texttt{qBound}: case of bound anaphors;
\item \texttt{recip}: reciprocal.
\end{itemize}

We introduce an XML-element \textless anaX\textgreater with its ID prefix ax as the root element for XML documents in the concrete XML annotation scheme CSyn\textsubscript{anaX} for coreference and other types of anaphoric link. Every annotation based on CSyn\textsubscript{anaX} is enclosed by a pair of the root element \textless anaX\textgreater and its closing element \textless /anaX\textgreater.

5 Illustrations

Here we introduce the first part of the dataset about a retired professor to show how it is annotated. The proposed CSyn\textsubscript{anaX} can annotate it, as shown below:

\begin{verbatim}
(50) a. Step 1: Identification of Markables
Not long ago {a retired professor}\textsubscript{41} of a small university\textsubscript{42} in Essex\textsubscript{14} lost {his\textsubscript{46} wealthy wife\textsubscript{27}}, who\textsubscript{45} had been ill for {over ten years}\textsubscript{29}. After {her\textsubscript{10} death\textsubscript{11}}, the professor\textsubscript{13} and {their\textsubscript{14} only daughter\textsubscript{15}}\textsubscript{16} found {a large sum of money}\textsubscript{17} left for them\textsubscript{18} to share.

b. Step 2a: Identification of Possible Anaphors and Step 2b: Links
\texttt{<anaX xml:id="ax1">}
\texttt{<entity xml:id="x1"}...
target="professor" type="person" form="name" definite="no" naturalGender="male"/>
<entity xml:id="x5" target="his" type="person" form="pro" naturalGender="male"/>
<entity xml:id="x6" target="wife" type="person" form="nom" definite="yes" naturalGender="female"/>
<entity xml:id="x8" target="who" type="person" form="pro"/>
<entity xml:id="x10" target="her" type="person" form="pro" naturalGender="female"/>
<entity xml:id="x13" target="professor" type="person" form="name" definite="yes" naturalGender="male"/>
<entity xml:id="x14" target="their" type="person" form="pro" plural="yes" naturalGender="male"/>
<entity xml:id="x15" target="daughter" type="person" form="nom" definite="yes" naturalGender="female"/>
<entity xml:id="x17" target="a large sum of money" type="impersonal" form="quant"/>
<entity xml:id="x18" target="them" type="personal" form="pro" plural="yes"/>
</anaX>

The pronoun their, is grounded on or anchored to the two antecedents, professor, and wife, whereas the pronoun them, is also grounded on the two antecedents, professor, and daughter. Note that professor, is grounded on the earlier occurrence of professor with an ID x1.

Consider one more fragment:

(53) a. Step 1: Identification of Markables:

b. Text:

Billy Jones, took {his family}, to {a manor}, in {the country}, near Edinburgh. They, spent most of a year, but stayed in {Cannes, France}, in the winter season.

The antecedent of the pronoun they, is split into Billy Jones, and family. The antecedent of the locative pronoun There, is ambiguous: it can be either a manor, or the country. This can be represented as in 54, for both of the situations hold:

(54) <anaLink xml:id="anaL3" figure="#x32" ground="#x29" relType="ident"/>
<anaLink xml:id="anaL3" figure="#x32" ground="#x30" relType="ident"/>

Consider how well-known donkey sentences are annotated. Two classical types are illustrated in (55):

(55) a. Every farmer who owns a donkey beats it.

b. If Pedro owns a donkey, he beats it.

Example (55a) can be annotated as below:

(56) a. Every farmer, who owns {a donkey}, beats it.

b. If Pedro owns a donkey, he beats it.

\[33\] Originally, from [22].
Example (55b) can be annotated as below:

(57) a. If Pedro owns {a donkey} to, he to beats it.

b. <anaX xml:id="ax4"> 
  <entity xml:id="x1" target="Pedro" type="person" form="name" /> 
  <entity xml:id="x2" target="a donkey" type="donkey" form="nom" /> 
  <entity xml:id="x3" target="he" type="person" form="pro" /> 
  <entity xml:id="x4" target="it" type="person" form="pro" /> 
  <anaLink xml:id="anaL1" figure="#x3" ground="#x1" type="ident" /> 
  <anaLink xml:id="anaL2" figure="#x4" ground="#x2" type="ident" /> 
</anaX>

Here both of the anaphor-antecedent or figure-ground pairs, <he, Pedro> and <it, a donkey> are treated as coreferring. Note that indefinite descriptions are treated as referential terms, not existential quantifiers (see [38].)

6 Semantic Interpretations: A Sketch

As stated earlier in the introductory section, every semantic annotation must be supported by an explicitly defined semantics. It is not something interesting or optional to do. Possibly unlike POS-tagging or syntactic parsing, the construction of a semantics for the abstract syntax of semantic annotation is a task required for any valid semantic annotation scheme as explicitly stipulated in [36]. An abstract syntax or a concrete syntax based on it alone cannot be justified without a supporting semantics.

The use of the lambda calculus in the line of [60] or that of the discourse representation structures (DRSs), proposed by [38] and [39], can, for instance, be linked to the abstract syntax to provide such a semantics for semantic annotations. Attempts have been made by [41], [68], [6, 8, 9], and [51] to develop an annotation-based semantics with the use of lambda calculus or by [7] with the use of DRSs. The use of lambda abstraction has run into the problem of complexity especially in dealing with multiple quantification and embedded adjunct structures. This should be the case with the treatment of various anaphoric phenomena. There are at least two interesting works to overcome this complexity problem: One is an earlier work by [62] which proposed a way of combining Montague semantics with DRSs and another is the most recent work by [9] which directly addresses to the treatment of anaphoric phenomena by combining underspecified representation (USR) that arises because of the presence of context-dependent expressions such as pronouns with annotation information representation (AIR). In constructing these representation structure, [9] shows how useful and necessary it is to combine the introduction of discourse referents in DRSs with markables in the annotation into USR and AIR, especially when there are multiple occurrences of identical anaphoric expressions, that is, pronouns, in a text.

Here is a very simple illustration showing how a list of annotations is converted into a semantic representation:

(58) a. Text: John hates himself.

b. Annotation:

<anaX xml:id="ax1" /> 
<entity xml:id="x1" target="John" type="name" form="name" /> 
<event xml:id="e1" target="hates" type="hate" /> 
<entity xml:id="x2" target="himself" type="person" form="pro" /> 
<naturalGender="male" /> 
<anaX xml:id="ax2" figure="#x2" ground="#x1" relType="ident" /> 
</anaX>

(59) Semantic Values

a. $\sigma_{\text{one}} := \lambda x[\text{name}(x, \text{John})]$  

b. $\sigma_{\text{two}} := \lambda y[\text{person}(y) \land \text{male}(y)]$ 

c. $\sigma_{\text{three}} := \lambda e[\text{hates}(e, x, y)]$  

d. $\sigma_{\text{anal}} := \lambda y(x = y)$ 

c. $\sigma_{\text{anal1}} := [\text{name}(x, \text{John}) \land \text{person}(y) \land \text{male}(y) \land \text{hates}(e, x, y) \land x = y]$ 

Or equivalently:

\[
x, y, e \\
\text{name}(x, \text{John}) \\
\text{person}(y) \\
\text{male}(y) \\
\text{hates}(e, x, y) \\
x = y
\]

(60) Semantic Values in DRT:

We leave detailed discussion of ways of interpreting anaphoric links as a work item for the future. Here are some remarks on the interpretation of various anaphoric expressions. First, names and definite descriptions are referential terms, both referring to some unique entities in the domain of discourse referents. Indefinite descriptions are also treated as referential terms, as mentioned earlier.

Second, as proposed and discussed in formal model-theoretic semantics (see [60], [3], [58], and [45]), proper names, definite descriptions, indefinite singular (a dog) or bare plural (donkeys) noun phrases as well as quantified noun phrases (three students, every man) are also interpreted as referring to sets of sets or properties, in the world. In our treatment, universally quantified expressions are differentiated from other types of generalized, but existentially quantified expressions.
Third, pronouns, on the other hand, do not refer directly to any entities in the world, but only through being coreferential with some other terms in the text (anaphoric use) or by referring to some entities that are provided contextually in a discourse situation (indexical use) (see [44]). Nevertheless, pronouns are also marked up as referring expressions or mentions in coreference annotation (see [16], [19], [20]).

There have been two trends in formal semantics that allow both of them to be treated as the same type of binding phenomena. One trend was initiated by [60] and further developed by [3], [58], [45], and a host of others. According to this trend, all noun phrases that include proper names (e.g., George Washington) and definite descriptions (e.g., the king of France) are treated as (generalized) quantifiers just like the universal and the existential quantifiers in logic and interpret them all as denoting sets of sets, as illustrated below:

\[
\begin{aligned}
&\alpha \in \text{universal quantifier} \\
&\exists \in \text{existential quantifier} \\
&\{X \mid X \in \alpha\} \\
&\{X \mid \exists X\} \\
&\{X \mid X \in \exists X\}
\end{aligned}
\]

Here, \([\alpha]^M\) stands for the denotation of an expression \(\alpha\) with respect to a model \(M\).

Possibly initiated by [40] among the linguists, another trend has been to treat pronouns as bound variables. As a variable, a pronoun \(x\) has its semantic \(\sigma(x)\) determined not by a model \(M\), but an assignment \(g\) that ranges over the domain \(D\) of a model \(M\) such that \(\sigma(x) = g(x)\), or written \([\sigma(x)]^g\). Since [40] there have been discussions, either in the framework of formal semantics or the binding theory of the generative grammar, on the status of pronouns among linguists and logicians such as [16], [29], [19], [20], [38], and [78]. As a result, various types of pronouns have come to be known, including pronouns, called ‘presumptive’, ‘intrusive’ or ‘lazy/paycheck’ pronouns. Here are some examples:

(61) a. \([\text{John}]^M = \{X \mid j \in X \land |X| = 1\}\)

b. \([\text{the dog}]^M = \{X \mid |\text{dog}|^M \cap X = 1\}\)

c. \([\text{all men}]^M = \{X \mid |\text{man}|^M \subseteq X\}\)

(62) Presumptive or intrusive pronoun:

a. This is the girl, that whenever it rains she cries.

b. This is the camel, that I think he likes Oscar.

(63) Lazy/paycheck pronoun:

c. John spent his paycheck on beer. Everyone else put it in the bank.

Another question has to do with the extension of markable mentions. Linguists such as [40], [29], [28], [42], [43], [88], [89] and too many others have discussed presumptive and lazy pronouns, while [74], [46], and other have discussed the occurrences of pronouns in modal contexts. Hence, to produce semantically adequate ASs, it may be necessary to accommodate all their ideas into a unified AS for anaphoric links.

The interpretation of indexical pronouns such as \(I\), \(we\) and you is more complex than their surface phenomena. It requires a clear understanding of dialog or discourse situations in which these indexicals interact with each other, especially in a language like Korean that has a complex honorific system. Little attention has, however, been given to the hearer-speaker mode exchange phenomena in communication except for [27] and many of his earlier works [25, 26]. [27] introduces the notion of anchoring that grounds the content of indexicals to real communication situations. In section 7, we introduce the elements such as \(<\text{metadata}\>\) and \(<\text{metaLink}\>\) to deal with the speaker and the hearer references that are often omitted in a pro-drop languages like Korean, but in a sketchy way to be able to understand such utterances:

(64) a. from [27], 11.2.2

John told Mary: I love you. Mary told John: I love you.

b. John to Mary: 사랑해, salang-hay. (I) love (you),

Mary to John: 사랑해. salang-hay. (I) love (you).

In the Korean examples, both the first-person pronoun and the second-person pronouns are deleted.

7 Application to Korean

7.1 Datasets

Our discussion of pronominalization in Korean starts with the following two versions of the Lord’s Prayer in Korean:

(65) Lord’s Prayer: Protestant Version

이름을 가리키는 (주관적인) 명사구는 반드시 보유어의 모집단이 불명사로부터 간접적으로 추정되는 관계를 가리키며, 이 관계가 중첩될 경우 발생할 수 있는 실례를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는これを 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를방지하기 위해, 한국어에서는 이를 방지하기 위해, ..., 한국어에서는 이를 방지하기 위해, 

(66) Lord’s Prayer: Catholic New Version

(67) a. from [27], 11.2.2

John told Mary: I love you. Mary told John: I love you.

b. John to Mary: 사랑해, salang-hay. (I) love (you),

Mary to John: 사랑해. salang-hay. (I) love (you).

In the Korean examples, both the first-person pronoun and the second-person pronouns are deleted.

7 Application to Korean

7.1 Datasets

Our discussion of pronominalization in Korean starts with the following two versions of the Lord’s Prayer in Korean:

(65) Lord’s Prayer: Protestant Version

이름을 가리키는 (주관적인) 명사구는 반드시 보유어의 모집단이 불명사로부터 간접적으로 추정되는 관계를 가리키며, 이 관계가 중첩될 경우 발생할 수 있는 실례를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를방지하기 위해, 한국어에서는 이를 방지하기 위해, ..., 한국어에서는 이를 방지하기 위해, 

(66) Lord’s Prayer: Catholic New Version

(67) a. from [27], 11.2.2

John told Mary: I love you. Mary told John: I love you.

b. John to Mary: 사랑해, salang-hay. (I) love (you),

Mary to John: 사랑해. salang-hay. (I) love (you).

In the Korean examples, both the first-person pronoun and the second-person pronouns are deleted.

7 Application to Korean

7.1 Datasets

Our discussion of pronominalization in Korean starts with the following two versions of the Lord’s Prayer in Korean:

(65) Lord’s Prayer: Protestant Version

이름을 가리키는 (주관적인) 명사구는 반드시 보유어의 모집단이 불명사로부터 간접적으로 추정되는 관계를 가리키며, 이 관계가 중첩될 경우 발생할 수 있는 실례를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를 방지하기 위해, 한국어에서는 이를방지하기 위해, 한국어에서는 이를 방지하기 위해, ..., 한국어에서는 이를 방지하기 위해, 

(66) Lord’s Prayer: Catholic New Version

(67) a. from [27], 11.2.2

John told Mary: I love you. Mary told John: I love you.

b. John to Mary: 사랑해, salang-hay. (I) love (you),

Mary to John: 사랑해. salang-hay. (I) love (you).

In the Korean examples, both the first-person pronoun and the second-person pronouns are deleted.
Both (65) and (66) show the rare use of pronouns in Korean: neither of them has any other pronoun than 우리 (wuli) corresponding to the pronoun our. The Protestant version of the Lord’s prayer has no word corresponding to thy or your, while the Catholic version has 아버지(apeci) ‘father’ for the addressee, thy in English. These datasets show the characteristics of Korean as a typical pro-drop language that allows the dropping not only of subject pronouns, but also of any other contextually recoverable pronouns. (67) are common examples of daily familiar greetings:

(67) a. 사람해요. (salang.hay.yo) ‘(I) love (you).’
   b. 고마워 잘 먹을게. (komawue. cal mek.ul.kkey) ‘(I) thank (you). (I) will eat (it) well’

The pro-dropping is often compensated by the repetition of the nouns referred to, as shown by (83) and a newly proposed Protestant version of the Lord’s prayer (68):

(68) 하늘에 계신 우리 아버지.
   아버지의 이름을 거룩하게 하시며
   아버지의 나라가 오게 하시며,
   아버지의 뜻이 하늘에서와 같이 맥에서 이루어지게 하소
   서.

Here are a couple of examples from [56] that show how noun phrases are repeated:

(69) 나는 감자가 많이 나는 곳장에서 태어나서 감자를 무척 좋아한다. 그리고 감자를 좋아 하는 사람을 만나면 해마다 감자를 적출로 보내고 살아간다.
   na-nun kamca-ka manh.i nanun kocang eyse thaye nase kamca-lul muchek coh.a hanta. kuliko kamca-lul coh.a hanun salam-ul mannymyen hay-mata kamca-lul senmul-lo ponayko siphecita.

‘Since I was born in a place where they grow a lot potatoes, I really like them. Moreover, whenever I meet someone who likes potatoes, I want to send them some as a present every year.’

(70) 나는 아버지를 사랑한다. 아버지가 없는 세상은 상상조차 할 수 없다. 아버지들은 정말 대지와도 같다.
   na-nun emeni-lul salang-hanta. emeni-ka epsnun seysang-un sangsang-cocha hal swu epsta. emeni.tul-un ceng-mal tayci-wato kaththa.

‘I love (my) mother. I can’t even imagine a world without her. Mothers are really like the earth to me.’

7.2 Production of Pronouns in Korean

[56] claim that there are no third-person pronouns in Korean. As an agglutinating language, Korean can, however, produce a host of third-person pronominal expressions, based on the three basic demonstrative pronouns, as listed in (71):

(71) a. 이 (i) ‘this’: pointing to some entity in close proximity
   b. 그 (ku) ‘that’: simply referring to some entity that has been mentioned or pointing to some entity in near proximity
   c. 저 (ce) ‘that (over there)’: pointing to some entity at a distance

[14] calls 이 (i) ‘this’ proximal, 그 (ku) ‘that’ medial, and 저 (ce) ‘that over there’ distal.

To these demonstrative pronouns, bound morphemes such as 문 (pun) ‘person’, 것 (kes) ‘thing’, or 는 (nom) ‘fellow, guy’ can be affixed to form pronouns. Here are examples:

(72) a. 이분이 제 선생님이시다.
   Ipun-i cey sensayingnim.i.sita
   This is my teacher.
   b. 내가 좋아하는 과일은 사과인데 배일 것을 하나씩 먹는다.
   nay-ka coh.a-hanun kwail-un sakwa.inte mayil kukes ul hanassik mek.nunta.
   ‘My favorite fruit is an apple and I eat it one a day.’
   c. 저놈은 도둑놈이다.
   cenom-un totwuknom.ita
   ‘That fellow over there is a thief.’

Bound nouns such as 때 (ttay) ‘time’ and 곳 (kos) ‘place’ can be affixed to those demonstratives:

(73) a. 내가 즐거던 그때, 내가 살던 곳을 영영 못 잊겠다.
   nay-ka culkiten kuttay, nay-ka salten kukos-ul yengyeng mos ic.keyssta
   ‘(I) can never forget the time I used to enjoy and the place where I used to live.’

7.3 Pragmatic Features Associated with Pronouns in Korean

The first and second person pronouns are seldom used in spoken language. If they are used, then their use is much constrained by pragmatic principles involving the speaker, the addressee(s), and the persons or things referred to in an utterance.
Table 1. Constraints on the Use of First and Second Person Pronouns

<table>
<thead>
<tr>
<th>pronouns</th>
<th>speechRole</th>
<th>plurality</th>
<th>speechStyle</th>
</tr>
</thead>
<tbody>
<tr>
<td>나나</td>
<td>speaker</td>
<td>no</td>
<td>plain</td>
</tr>
<tr>
<td>저저</td>
<td>speaker</td>
<td>no</td>
<td>humble</td>
</tr>
<tr>
<td>우리나라</td>
<td>speaker</td>
<td>yes</td>
<td>plain</td>
</tr>
<tr>
<td>저저</td>
<td>addresssee</td>
<td>no</td>
<td>plain, honorific</td>
</tr>
<tr>
<td>당신당신</td>
<td>addresssee</td>
<td>no</td>
<td>plain</td>
</tr>
<tr>
<td>너희니</td>
<td>addresssee</td>
<td>yes</td>
<td>plain</td>
</tr>
<tr>
<td>여러분여러분</td>
<td>addresssee</td>
<td>yes</td>
<td>plain</td>
</tr>
</tbody>
</table>

나 (na), 저 (ce), 우리 (wuli), and 저희 (cehui) are first-person pronouns, each referring to a speaker or speakers. 나 (na) is a plain speech style form, while 저 (ce) is called a humble or self-lowering speech style form, meaning that the speaker is considered as paying due respect to the addressee(s). 우리 (wuli) and 저희 (cehui) are plural forms, corresponding to the singular forms 나 and 저, respectively. 38저저 (cehui) is a humble form. 

너 (ne), 자네 (caney), 당신 (tangsin), 너희 (nehui), and 여러분 (yelepun) are second-person pronouns, referring to an addressee or a group of addressees. The first two forms, 너 (ne) and 자네 (caney), and the two plural forms 너희 (nehui) and 여러분 (yelepun) are of the plain speech style form, whereas the other three 당신 (tangsin) is an addressee-honoring form or plain form. The use of these three forms is very much restricted, too. In prayers or some writings, 당신 (tangsin) is accepted as a humble form, but in ordinary dialog situations is often misunderstood as failing to convey respect to the addressee. There are many other addressee-referring forms such as 어른 (elusin) ‘elder’, 부인 (puin) ‘lady’, 선생님 (seonsayingnim) ‘teacher’, 사모님 (samomon) ‘teacher’s wife’, 강아지 (kang.aci) ‘puppy’, 삼촌 (samchwon) ‘uncle’, 이모 (imo) ‘aunt’, 오빠 (oppa) ‘brother’, 자기 (caki) ‘self’, but their use is very much pragmatically constrained, reflecting various social-level relations between the speaker and the addressee. 어른 (elusin) is used only when an elderly person is addressed to, but older people may not want to be called as such. 자기 (caki) is a third-person reflexive pronoun, but is often used among young married couples, addressing to each other.

By suffixing a plural marker 들 (tul), we can have the plural forms, 저희들 (cehui-tul) and 너희들 (nehui-tul). Other such plural forms are: 당신들 (tangsin-tul) and 여러분들 (yelepun-tul).

When followed by a nominative (subject) marker, 나 (na), 저 (ce), and 너 (ne) undergo a vowel-raising process, thus changing their forms to 내 (nay), 계 (cey), and 네 (ney), respectively. These forms are also used as genitive case forms. 

(74) a. 저를 용서하여 주십시오. 제가 잘못 했습니다.
   ce-lul yongse-haye cwuisipcyo. cey ka calpos haysswupnita.
   ‘Pardon me. I did wrong.’

The use of 자기 (caki) is restricted. It cannot be used with the first person pronoun, although 자신 (casin) is acceptable.

(76) a. 내 자신 것을 몰랐다.
   nay casin ku.kes-ul moll.assta.
   ‘I myself didn’t know it.’

b. 미야가 자기 자신들을 자랑했다.
   mia-ka caki casik.tul-ul calang.hayssta
   ‘Mia boasted about her own children.’

The second person noun 당신 (tangsin) can be used as a third person reflexive, as in (77):

(77) 할아버지가 당신 손주들을 무척 아끼신다.
   halapeci-ka tangsin soncwu.tul-ul mwuchek akki.sinta
   ‘Grand is very fond of his own grandchildren.’

Here are examples of reciprocals in Korean:

(78) a. 서로 (selo) ‘each other, one another’
   너희들은 서로 사랑하다.
   nehui.tul-un seolo salang.hayela
   ‘You love each other.’

b. 서로로서 (seloselo) ‘each other, one another’
   우리들은 서로로서 손을 맞잡고 노래 부르자.
   wuli.tul-un seloselo son-ul mac.cap.ko nolay pwulu.ca
   ‘Let’s hold hands with each other and sing a song.’

There are two forms of quantifiers: one is an adnominal form and the other an adverbial form, as shown in (79):

(79) a. Adnominal Form:
   모든 (motun) ‘all’, 각 (kak) ‘each’
   모든 사람이 모였다. 각각 손에 태극기를 들고
   motun salam-i moy.essta. kaka son-ey thaykuk.ki-lul
tul.ko
   ‘All people gathered, each holding a Korean flag in their hand.’

b. Adverbial Form:
   모두 (motwu) ‘all’, 각각 (kak.kak) ‘each’
   사람들 모두 모였더. 그들은 각각 손에 태극기를 들고서
   salam.tul-i motwu moy.essta. ku.tul-un kakkak son-ey
   thaykukki-lul tul.ko.se
   ‘People all gathered, each holding a Korean flag in their hand.’
7.5 Extension of ASyn\textsubscript{ana} to Korean

7.5.1 General

As in English and other languages, the reference of first-person and the second-person pronouns is determined by each discourse situation. First-person pronouns refer to the speaker, sender or creator of a piece of information, while second-person pronouns refer to the addressee or recipient of that information.

To construct any pragmatically relevant relationship between the speaker and the addressee, we also introduce the discourse situation or circumstance where a particular text is created. It is necessary to devise a way of constructing such a situation or circumstance.

7.5.2 Discourse Situation as Part of Metadata

To construct a discourse situation, we introduce at least four entity types that constitute it.

(80) 내가 너를 사랑하는데 너도 나를 사랑하시는가? 우리가 서로 사랑하는지?
    nay-ka ne-lul salang.ha.nunte ne-to na-lul salang.ha.nunya
    wuli-ka selo salang.ha.nunci?
    ‘I love you. Do you, also love me? Do we, also love each other,?’

Just like any other indexical expressions such as now and here, the first-person and the second-person pronouns are also indexical expressions, depending the situation in which a discourse is carried out or the circumstance where a particular text is created. It is necessary to devise a way of constructing such a situation or circumstance.

The feature assignment @ for each of the four entity types is specified in (82):

(82) a. Feature Assignment @\textsubscript{speaker}:
    attributes = identifier, target, [plural], [speechStyle], [comment];
    identifier = idPrefix="sp" followed by a positive integer;
    target = IDREF|CDATA;
    plural = "yes";
    {*If not plural, unspecified and the value is assumed to be singular.*}
    speechStyle ="humble";{*If not humble, it is a plain style.*}
    comment = CDATA;

    b. Feature Assignment @\textsubscript{addressee}:
    attributes = identifier, target, type, plurality, [speechStyle], [comment];
    identifier = idPrefix="ad" followed by a positive integer;
    target = IDREF|CDATA;
    speechStyle = "humble";{*If not humble, it is a plain style.*}
    comment = CDATA;

    c. Feature Assignment @\textsubscript{uTime}:
    attributes = identifier, target, value, comment;
    identifier = idPrefix="ut" followed by a positive integer;
    target = IDREF|CDATA;
    value = CDATA;{*follows [32].*}
    comment = CDATA;

    d. Feature Assignment @\textsubscript{uPlace}:
    attributes = identifier, target, value, comment;
    identifier = idPrefix="up" followed by a positive integer;
    target = IDREF|CDATA;
    value = CDATA;{*may follow [35].*}
    comment = CDATA;

    e. Feature Assignment @\textsubscript{metaLink}:
    attributes = identifier, figure, ground, relType, [trigger], comment;
    identifier = idPrefix="metaL" followed by a positive integer;
    figure = IDREF;
    ground = IDREFS;
    {*The value of the attribute @ground is either a speaker or an addressee. It can also have more than one value.*}
    relType = CDATA;
    trigger = IDREF;
    comment = CDATA;

The attribute names @figure and ground are the terms borrowed from Gestalt psychology, as used extensively for the annotation of spatial relations in [35]. The figure is a perceptually prominent object against some background which constitutes its ground. The attribute speechStyle is relevant only for those languages like Korean and Japanese that have a honorific system of discourses.

In order to incorporate the metadata information on discourse situations into the anaphoric annotation scheme, specified with ASyn\textsubscript{ana} and CSyn\textsubscript{anaX}, we need to instantiate ASyn\textsubscript{semAnn} with an XML-based CSyn\textsubscript{semAnnX} from the general annotation scheme specified with ASyn. The concrete syntax thus extended will have <\texttt{semAnn}> with an ID prefix\textsubscript{ana} as its root element under which the two elements <\texttt{metaData}> and <\texttt{anaX}> are embedded.
7.5.3 Illustrations

Consider the Catholic new version of the Lord’s prayer.:

(83) Lord’s Prayer: Catholic New Version

This is the prayer that Jesus taught his disciples. The annotation of this prayer undergoes two annotation processes: (a) word segmentation, (b) metadata, (c) identification of markables and (d) the annotation of links.

(84) a. Word Segmentation:

(b) Consider example (85), which is identical with (70):

(85) 나는 어머니를 사랑한다. 어머니가 없는 세상은 상상조차 할 수 없다. 어머니들은 정말 대지와도 같다.

There are three occurrences of 어머니 (emeni) ‘mother’, but the third one 어머니들 (emeni.tul) ‘mothers’ is a plural indefinite form.

(86) shows how (85) is annotated. Note that the fragment has undergone word segmentation as a preprocessing step:

b. <semAnn xml:id="sa2">

As a final example, consider example (ex70):

(87) 내가 너를 사랑하는데 너도 나를 사랑하느냐? 우리가 서로 사랑하는지?

‘I, love you.$_i$. Do you$_j$ also love me$_i$? Do we$_i,j$ love each other$_i,j$?’

This example contains first-person and second-person pronouns. Their reference is determined by a discourse situation, as shown in (88b) Metadata:

(88) a. Word Segmentation:

This example contains first-person and second-person pronouns. Their reference is determined by a discourse situation, as shown in (88b) Metadata:

b. Metadata:

In order to interpret reciprocal$(A,R)$, we adopt the definition of Weak Reciprocity (WR), discussed by [50]:

(90) Definition of Weak Reciprocity:

Given a set $A$ and a relation $R$, which is $A \times A$, reciprocal$(A,R)$ is true if and only $\forall x \in A \exists \{y,z\} \in A [x \neq y \land x \neq z \land A R y \land A R z]$

reciprocal$(A,R)$ in (89) is now interpreted as in (91):

(91) $[\text{speaker}(x) \land \text{addressee}(y) \land \text{love}(e_1,x,y) \land \text{love}(e_2,y,x)]$

The interpretation (91) is a case of symmetric reciprocity. It is a particular case of (90) in which the cardinality of $A$ is 2, $|A| = 2$, such that $z = y$.

8 Concluding Remarks: A Summary

This paper is an extended and revised version of [54]. It mainly aimed at making some linguistic contributions as necessary ground work to an ISO initiative to produce an international standard on anaphoric annotation as part of a series of semantic annotation schemes. It has thus reviewed several existing annotation schemes for anaphoric links including TEI-based [5], [85], [30], [79] based on [61], [64], [65], and [70] in efforts to normalize them into an interoperable annotation scheme implementing common denominators as an ISO international standard on semantic annotation.

Section 4 through section 6 form the core of this paper. Following the initial proposal of [7] and his subsequent works including [10], a basic work for [36], this paper adopts a formal approach to the construction of a semantic annotation scheme for anaphoric annotation which consists of:

1. an abstract syntax $ASyn_{ana}$,
2. a set of possible concrete syntaxes $CSyn_{ana}$, and
3. a semantics supporting the abstract syntax.

As designed by [55], the proposed abstract syntax $ASyn_{ana}$ for anaphoric annotation is derivable from the general abstract syntax $ASyn$, providing an anaphoric annotation scheme structured with a two-level specification language in abstract terms:

1. the level of identifying entities, called ‘discourse entities’, as marakbles and
2. the level of linking each anaphor to one or more antecedents, called ‘anchors’ or ‘grounds’. $ASyn_{ana}$ allows a variety of types of anaphoric linking for coreference and other types of anaphoric relations by simply specifying their data type only.

\[40\] The Korean reciprocal 서로 (selo) is a one single word. It cannot be subject to the analysis of [21] and [18] that relates each other to each the other in terms of a distributive quantifier each and some other elements such as coordinating elements of distributive adverbs.
ASyn\textsubscript{ana} allows the derivation of a class of concrete syntaxes CSyn\textsubscript{anaX} that are semantically equivalent, but each with a different representation format. The proposed concrete syntax CSyn\textsubscript{anaX} is based on XML, providing an XML-based representation format. The basic entity type for discourse entities that are referred to by either referring or anaphoric expressions is tagged <entity/> with its ID prefix being “X” which is followed by an integer for unique identification. The attribute-value specification associated with <entity/> has a long list of optional or implied attributes such as @definite, @naturalGender and @plurality other than a small number of required attributes, @xml:id, @target, @type and @form, depending on the language to which the annotation scheme is applied.

The XML-based concrete syntax CSyn\textsubscript{anaX} for English has a single linking tag <anaLink/> for anaphoric relations. Besides its ID with its prefix "ana", which is needed for doing semantics as a handle for combinatorial operations, the element <anaLink/> has three required attributes: @figure, @ground, and @relType. The attribute @ground may have more than one antecedent as its value. The list of values for the attribute @relType includes: "ident" for coreference, "partIdent" for partial identity, "splitIdent" for partial identification, and "recip" for reciprocals.

The formulation of a semantics for an abstract syntax is required if the abstract syntax is for semantic annotation. Section 6 thus showed how semantic forms can be derived from the list of annotations or XML-elements in CSyn\textsubscript{anaX} in a compositional manner in order to validate the proposal CSyn\textsubscript{ana}, even partially. The task of constructing a model-theoretic semantics for CSyn\textsubscript{ana} is left as a future agenda item.

By applying ASyn\textsubscript{ana} to Korean, we have shown how adequately a concrete syntax CSyn\textsubscript{anaX} can be modified and extended. One such an example is the treatment of the reciprocal expression 서로 (selo) ‘each other’ in Korean. Given an adequate semantics of reciprocity in language, reciprocal expressions in Korean as well as in English are shown to be adequately annotated and interpreted.

In treating Korean, we have also extended ASyn\textsubscript{ana} to accommodate the notion of discourse situations as metadata, tagged <metaData/>, that specifies various discourse-involving entities such as the speaker, the addressee(s), the utterance time and the utterance place as well as a link that relates them with each other. These entities are then tagged in XML: <speaker>, <addresse>, <uTime>, and <uPlace> with the link being tagged <metaLink/>. The element, tagged <metaData/>, is the root element for the annotation of discourse situations.

The current versions of proposed ASyn\textsubscript{ana} and CSyn\textsubscript{anaX} both need to be evaluated against inter-annotator agreed annotations with large data, especially applied to Korean. Such task is, however, left as a future project.

Acknowledgments

I owe thanks to Harry Bunt for his great suggestions and also to Suk-Jin Chang, Jae-Woong Choe, Roland Hausser, Hwan-Mook Lee, Do-sam Hwang, and four anonymous reviewers for reading the original version of this paper submitted to the 2014 ISA-10 Workshop, LREC 2014, with encouraging comments. I also thank for the reviewers of this extended paper whose comments I have benefited much for its improvement. Thanks to Roland Hausser and Jae-Woong Choe again for reading some crucial parts of the pre-final version of this paper.

Notes

\footnote{This paper is an extended and revised version of a paper with the same title that appeared in Harry Bunt (ed.), Proceedings of the 10th Joint ISO-ACL SIGSE Workshop on Interoperable Semantic Annotation pp.29–38, a satellite workshop ISA-10 of LREC 2014, May 26, 2014, Reykjavik, Iceland.}

REFERENCES


