OntoNotes English Co-reference Guidelines

Version 7.0

1 Overview 3
1.1 Mentions 3
1.1.1 Noun phrases 3
1.1.2 Possessives 3
1.1.3 Premodifiers 3
1.1.4 Verbs 3
1.2 Co-reference link types 4
1.2.1 Identical (IDENT) 4
1.2.2 Appositives (APPOS) 4
2 IDENT (anaphoric co-reference) 4
2.1 Pronouns and Demonstratives 4
2.2 Generic Mentions 5
2.3 Pre-modifiers 7
2.4 Nested Mentions 8
2.4.1 Head-Sharing NPs 8
2.4.2 Proper Names 9
2.5 Copular Structures 9
2.6 Determining which entity to add 9
2.7 Small clauses 10
2.8 Temporal expressions 10
3 APPOS (appositives) 10
3.1 Marking appositive heads 11
3.2 Linking appositive spans to other referents 11
3.3 Special cases 12
4 Special Issues 13
4.1 Organization and members 13
4.2 Gender and Number 13
4.3 Indefinite uses of proper nouns 13
4.4 GPEs and governments 14
4.5 Quantifying Expressions 14
4.5.1 Quantifiers 14
4.5.2 Partitives 14
4.5.3 Linking quantifying expressions to other mentions 15
4.6 Possessive extents 16
4.7 Formulaic mentions 16
4.8 Sentence fragments 16
4.9 Metonyms 16

Supplement A 18
Supplement B 20
Supplement C 25
Supplement D 27
Supplement E 29

1 Overview

According to the guidelines written for a name-tagging task at Georgetown University, "A human reading [a text] is able to understand it using her knowledge of language as well as her knowledge of the world. To get a computer to do the same, it is helpful to prepare examples of text marked up with whatever information the human needed to extract from it. The resulting corpus of annotated examples can then be used to teach the computer to [automatically] extract the same kind of information." (Georgetown Guidelines for Protein Name-Tagging, 2004)

The purpose of this OntoNotes task is to co-reference, or 'link,' all the specific mentions in a text that point ('refer') to the same entities and events, and to distinguish between types of co-reference as needed to improve accuracy and scope. Texts annotated in this way will help the computer learn to correctly identify multiple mentions of the same entity. Co-reference is limited to noun phrases (NPs), possessives, proper noun premodifiers (PreMods), and verbs. This initial overview describes the types of mentions and the types of co-reference applied.

1.1 Mentions

1.1.1 Noun phrases

A noun phrase consists of a noun or pronoun and its modifiers. All noun phrases with distinct headwords are extracted from previously treebanked data and presented to annotators as highlighted spans prior to annotation. Except in specific cases, annotators do **not** need to identify or manually extract any additional NP spans. Whenever head-sharing NPs are nested, the largest logical span is used in co-reference (see 2.4.1).

1.1.2 Possessives

Possessive nouns should be co-referenced to other mentions. Possessive proper nouns *(Fred's)* are extracted from the treebanked data; however, possessive pronouns *(his)* must be manually extracted by the annotator and added to the list of mentions:

(1) [Fred's]$_x$ wife is Wilma, and [his]$_x$ daughter is Pebbles.

1.1.3 Premodifiers

A premodifier (PreMod) is a word that precedes and modifies a noun. Proper noun PreMods can be co-referenced to existing noun phrases **and/or** other proper PreMods, and should be manually extracted by the annotator and added to the list of mentions. Non-proper and adjectival premodifiers are **not** eligible for co-reference (see 2.3).

1.1.4 Verbs

Verbs can be manually extracted by the annotator and added as single-word spans **if** (and only if) they can be co-referenced with an existing noun phrase. This includes morphologically related nominalizations, as in (2), and noun phrases that refer to the same event but are lexically distinct, as in (3).

(2) Sales of passenger cars [grew]$_x$ 22%. [The strong growth]$_x$ followed year-to-year increases.

(3) Japan's domestic sales of cars, trucks and buses in October [rose]$_x$ 18% from a year earlier to 500,004 units, a record for

the month, the Japan Automobile Dealers' Association said. [The strong growth]$_x$ followed year-to-year increases of 21% in August and 12% in September.

**Only the single-word head of the verb phrase is included in the span, even in cases where the entire verb phrase is the logical co-referent.

1.2 Co-reference link types

Two types of co-reference chains are marked: Identical (IDENT) and Appositive (APPOS).

1.2.1 Identical (IDENT)

Names, nominal mentions, pronominal mentions, and verbal mentions of the same entity, concept, or event are co-referenced as IDENT. There is no restriction on which semantic types can be considered for co-reference; in particular, co-reference is **not** limited to ACE types.

(4) She had [a good suggestion]$_x$, and [it]$_x$ was unanimously accepted.

- IDENT chain: [a good suggestion], [it]

1.2.2 Appositives (APPOS)

Appositive constructions consist of two (or more) immediately-adjacent noun phrases, separated only by a comma, colon, dash, or parenthesis. The first NP is the head, or referent, which points to a specific object/concept in the world, and the adjacent NP(s) specify one or more attributes of that referent, thus renaming or further defining the head.

(5) [[John]$_x$, [a linguist]$_x$], is coming to dinner.

- APPOS chain: [John]$_{HEAD}$, [a linguist]$_{ATTRIB}$

2 IDENT (anaphoric co-reference)

The IDENT type is used for anaphoric co-reference, meaning links between pronominal, definite nominal, and proper nominal (named) mentions of specific referents.

(6) [Elco Industries Inc.]$_x$ said [it]$_x$ expects net income in the year ending June 30, 1990, to fall below a recent analyst's estimate of $ 1.65 a share. [The Rockford, Il. maker of fasteners]$_x$ also said...

- [Elco Industries Inc.](proper nominal), [it](pronominal), [The Rockford, Il. maker of fasteners](definite nominal)

Anaphoric co-reference does not include entities that are only mentioned as generic, underspecified or abstract (See 2.2).

2.1 Pronouns and Demonstratives

All pronouns and demonstratives are linked to their referents, even if they occur in quoted speech, as in example (7), or are nested within the span of the referent, as in example (8):

(7) Although [Mr. Clinton]$_x$ is out of office, [he]$_x$ says [he]'ll still be around. "[I]$_x$ left the White House, but [I]$_x$'m still here."

- IDENT chain: [Mr. Clinton], [he], [he], [I], [I]

(8) The company plans to market [a single-engine plane]$_x$ with a parachute for [the plane [itself]$_x$]$_x$

- IDENT chain: [a single-engine plane], [the plane itself], [itself]

Possessive pronouns are always linked to their antecedents, and must be manually extracted by the annotator.

(9) [Fred]$_x$ is married to Wilma, and [his]$_x$ daughter is Pebbles.

- IDENT chain: [Fred], [his](manually extracted)

Possessive adjectives (*yours, mine, ours, etc.*), however, are pre-extracted by treebank. These should be linked to the **possessor**, even though they also contain an implicit reference to the item possessed:

(10) This bench is [mine]$_x$. [I]$_x$ sit here every day.

- IDENT chain: [mine], [I]

Instances of the generic *you* are **not** linked. In the following example, **none** of the four occurrences of the pronoun "you" are eligible for coreference:

(11) A lot of times, Frist recalls, [you]'d have a critical patient lying there waiting for a new heart, and [you]'d want to cut, but [you] couldn't start unless [you] knew that the replacement heart would make it to the operating room.

Pronouns used as expletives (sometimes called 'dummy' pronouns) are also **not** linked. These pronouns do not refer to any specific entity, and so cannot be linked to another noun phrase. In the example below, 'it' and 'there' are expletive pronouns, and should remain unlinked.

(12) Since [there] have been no further negotiations, [it] seems obvious that the violence will continue.

As a test for expletive pronouns, remember that normal pronouns can be replaced with a noun phrase, while expletive pronouns cannot.

2.2 Generic Mentions

Generic, underspecified, and abstract nominal mentions are linked to referring pronouns and definite mentions of the same entity, but **not** to other generic nominal mentions.

Bare plurals, such as "officials" in (13), "meetings" in (14), and "parents" in (15) are always generic. As such, they can form an IDENT chain with any subsequent non-generic mentions.

However, because generic mentions cannot be linked to one another, a new IDENT chain is required for each generic mention in example (15).

(13) [Officials]$_x$ said [they]$_x$ are tired of making the same statements.

- IDENT chain: [Officials](generic), [they](pronoun)

(14) [Meetings]$_x$ are most productive when [they]$_x$ are held in the morning. [Those meetings]$_x$, however, generally have the worst attendance.

- IDENT chain: [Meetings](generic), [they](pronoun), [Those meetings](definite)

(15) [Parents]$_x$ should be involved with their children's education at home, not in school. [They]$_x$ should see to it that [their]$_x$ kids don't play truant; [they]$_x$ should make certain that the children spend enough time doing homework; [they]$_x$ should scrutinize the report card. [Parents]$_y$ are too likely to blame schools for the educational limitations of [their]$_y$ children. If [parents]$_z$ are dissatisfied with a school, [they]$_z$ should have the option of switching to another.

- IDENT chain X: [Parents](generic), [They](pronoun), [their](pronoun), [they](pronoun), [they](pronoun)
- IDENT chain Y: [Parents](generic), [their](pronoun)
- IDENT chain Z: [Parents](generic), [they](pronoun)

Indefinite noun phrases, which begin with the indefinite article (*a, an*), are also considered generic. In (16) "an agreement" **cannot** be linked to "a final deal" since both NPs are indefinite.

(16) Israeli-Palestinian peace talks have ended in Egypt with a statement declaring the two sides are closer than ever to [an agreement]..."We are closer than ever to the possibility of striking [a final deal] between us and the Palestinians."

Abstract and underspecified nominal mentions similarly cannot be linked to one another.

(17) Allergan Inc. said it received approval to sell the PhacoFlex intraocular lens, the first foldable silicone lens available for [cataract surgery]. The lens' foldability enables it to be inserted in smaller incisions than are now possible for [cataract surgery].

One exception to the rules of generic mentions occurs when a news anchor offers a brief overview, similar to a headline, before discussing the details of a story. This often results in two sequential generic mentions of the entities involved: once in the introduction and once in the story itself. In this instance alone, the entities in the introduction **should** be linked to the corresponding mentions within the story, even though both are technically generic:

(18) "Next, [a man]$_x$ robs [a bank]$_y$ at gunpoint, but leaves [[his]$_x$ wallet]$_z$ behind. Late yesterday afternoon, [a man]$_x$ posing as a customer demanded money from the teller at [a local Bank of America branch]$_y$. [The suspect]$_x$ was later apprehended at home by police, who found [his]$_x$ name and address in [a wallet]$_z$ recovered at the scene."

- IDENT chain X: [a man](indefinite-intro), [his](pronoun), [a man](indefinite-story), [The suspect](definite), [his](pronoun)
- IDENT chain Y: [a bank](indefinite-intro), [a local Bank of America branch](indefinite-story)
- IDENT chain Z: [his wallet](definite), [a wallet](indefinite-story)

2.3 Pre-modifiers

Premodifiers must be proper nouns in order to be manually extracted for co-reference. Adjectives and non-proper nouns **cannot** be linked as PreMods, even if they seem to refer to other mentions of the same entity, as in (20) and (21).

Note that **only the premodifying noun itself is included in the PreMod span, since any preceding articles (*the, a, an*) belong to the full noun phrase.

(19) But [the Army Corps of Engineers]$_x$ expects the river level to continue falling this month. "The flow of the Missouri River is slowed," an [Army Corps]$_x$ spokesman said.

- IDENT chain: [the Army Corps of Engineers], [Army Corps](proper PreMod, manually extracted)

(20) [Wheat] is an important part of the economy in the Midwest. In Kansas, wheat fields stretch as far as the eye can see.

- *wheat* fields – non-proper premodifier (no coref.)

(21) [Charles Dickens] was famous for his memorable characters. The Dickensian character has since become a literary archetype.

- *Dickensian* character – adjectival premodifier (no coref.)

Pre-modifying dates and monetary amounts are also eligible for co-reference.

(22) The current account deficit on France's balance of payments narrowed to 1.48 billion French francs in August from a revised 2.1 billion francs in [July]$_x$, the Finance Ministry said. Previously, the [July]$_x$ figure was estimated at a deficit of 613 million francs.

- IDENT chain: [July], [July](date PreMod)

(23) The company's [$150]$_x$ offer was unexpected. The firm balked at [the price]$_x$.

- IDENT chain: [$150](monetary PreMod), [the price]

Acronymic premodifiers should be co-referenced unless they refer to nationality (see example (29) below). In the examples (24) and (25), "FBI" and "U.N." are eligible for co-reference.

(24) the [FBI] spokesman

(25) the [U.N.] Secretary General

Nationality acronyms and other adjectival forms of GPEs, however, are **not** eligible for co-reference as premodifiers. (Although nationality acronyms can always occur as proper noun phrases, as in (26) below.) Thus, only example (27) below contains a linkable PreMod.

(26) relations between [the U.S.] and Japan – proper noun phrase

(27) the [United States] policy – proper noun PreMod

(28) the American policy – nationality adjective (no coref.)

(29) the U.S. policy – nationality acronym (no coref.)

Even when nationality acronyms act like their non-acronymic counterparts, they are **not** considered proper premodifiers. In example (30) "Japan" can be co-referenced as a PreMod, but "U.S." **cannot**:

(30) On U.S.-[Japan] relations: "I'm encouraged."

Proper noun pre-modifiers that include acronyms in the span, however, **are** eligible for co-reference:

(31) A [U.S. Treasury] spokesman

2.4 Nested Mentions

2.4.1 Head-Sharing NPs

Head-sharing NPs are two (or more) extracted entities, the shorter one(s) contained within the span of the longer, sharing the same content word as their headword. In such cases, the longest logical span should be used in co-reference with other mentions.

(32) There's already word of [[[a possible Israeli-Palestinian **summit**] in Egypt] in the next several days]$_x$. [This summit]$_x$ ...

- IDENT chain: [a possible Israeli-Palestinian summit in Egypt in the next several days], [This summit]

2.4.2 Proper Names

Proper names, including the titles of songs and other works of art are considered to be atomic, and nested mentions inside proper names are **not** annotated separately. In the following examples, the location names that form part of the organization names are **not** eligible for co-reference.

(33) [*Massachusetts* Institute of Technology]

(34) [Bank of *America*]

(35) [the *Chicago* Board of Trade]

(36) ["I Left My Heart in *San Francisco*"]

2.5 Copular Structures

A copular structure consists of a referent (usually the subject), an attribute of that referent (usually the predicate), and a copula (most often, though not always, a 'linking' verb). The copula serves to equate (or link) the referent with the attribute. Relationships signaled by copular structures will be captured through word sense tagging, and annotators should **not** mark co-reference between the two elements:

(37) [John] is [a linguist] (no co-ref.)

Some common copular verbs are: *be, appear, feel, look, seem, remain, stay, become, end up, get*. In the following example, no co-reference is marked between mentions, since "called" is copular.

(38) Called [Otto's Original Oat Bran Beer], [the brew] costs about $12.75 a case.

Not all copular structures include a verb. In this example, "or" functions as a copula; therefore, neither an IDENT nor an APPOS relation is marked.

(39) Among other things, Mr. Bologna said that the sale will facilitate Gen-Probe's marketing of a diagnostic test for [acquired immune deficiency syndrome], or [AIDS].

2.6 Determining which entity to add

Only the leftmost element of a copular structure (the subject, or referent) should be linked to any subsequent mentions. The predicate, or attribute, remains unlinked.

(40) $[John]_x$ is a [linguist]. $[People]_y$ are nervous around $[John]_x$, because $[he]_x$ always corrects $[their]_y$ grammar.

- IDENT chain X: [John], [John], [he]
- IDENT chain Y: [People], [their]

In the example below, "John Smith" is the attribute, and so remains unlinked, while "the president of the bank" is the referent, and is linked to other mentions.

(41) [The president of the bank]$_x$ is [John Smith]. [He]$_x$ loves [his]$_x$ job.

- IDENT chain: [The president of the bank], [He], [his]

2.7 Small clauses

(42) John considers [Fred] [an idiot].

"Fred" and "an idiot" are **not** linked. This small clause construction is interpreted as missing the copula ("John considers Fred *to be* an idiot").

2.8 Temporal expressions

Temporal expressions are eligible for co-reference, including deictic expressions such as: *now, then*, *today, tomorrow, yesterday,* etc. or other temporal expressions that are relative to the time of the writing of the article.

(43) John spent [three years]$_x$ in jail. In [that time]$_x$...

- IDENT chain: [three years], [that time]

(44) The limit could range [from three years to seven years]$_x$, depending on the composition of the management team and the nature of its strategic plan. At [the end of [this period]$_x$]$_y$, the poison pill would be eliminated automatically, unless a new poison pill were approved by the then-current shareholders, who would have an opportunity to evaluate the corporation's strategy and management team at [that time]$_y$.

- IDENT chain X: [from three to seven years], [this period]

- IDENT chain Y: [the end of this period], [that time]

Multi-date temporal expressions (e.g. *month, day, year*), are considered atomic, and nested dates are **not** linked to other mentions of that date. In (45) below, there can be no co-ref chains for "November" or "2000" because "November 7, 2000" is atomic.

(45) American presidential elections are always held in [November]. The 2000 election was held on [November 7, 2000]$_x$. However, a winner was not declared on [that day]$_x$.

- IDENT chain: [November 7, 2000], [that day]

3 APPOS (appositives)

An appositive construction contains a noun phrase that is modified by one or more immediately-adjacent noun phrase(s), which are separated by only a comma, colon, or parenthesis. APPOS chains consist of a HEAD, or referent (a noun phrase that points to an object or concept in the world), and one or more ATTRIBUTES, which rename or further define that referent.

(46) [John]$_{x\text{-HEAD}}$,[a linguist I know]$_{x\text{-ATTRIB}}$, is coming to dinner.

**As a test for apposition, either part of the appositive by itself should make sense with the rest of the sentence.

- [John] is coming to dinner.
- [A linguist I know] is coming to dinner.

3.1 Marking appositive heads

For each appositive construction, the most specific element is marked as the head. According to the specificity scale below, the head of the APPOS chain in (46) above should be "John" (a proper noun) because it is more specific than "a linguist" (an indefinite NP).

(MOST SPECIFIC)----------(LEAST SPECIFIC)
proper noun > pronoun > definite NP > indefinite specific NP > non-specific NP
John > he > the linguist > a linguist I know > noted linguist

In the examples below, the underlined element is the most specific, and therefore is marked as the head of the construction

(47) [John Smith]$_{x\text{-}HEAD}$, [noted linguist]$_{x\text{-}ATTRIB}$,

(48) [A famous linguist]$_{x\text{-}ATTRIB}$, [he]$_{x\text{-}HEAD}$ studied at MIT

(49) [the president of the linguistics club]$_{x\text{-}ATTRIB}$, [J. Smith]$_{x\text{-}HEAD}$

In cases where the two members of the APPOS chain are equivalent in specificity, as in (50), (51), and (52), the left-most member of the appositive will be considered the head.

**Recall that definite NPs are those preceded by a definite article or a possessive.

(50) [The chairman]$_{x\text{-}HEAD}$, [the man who never gives up]$_{x\text{-}ATTRIB}$

(51) [The sheriff]$_{x\text{-}HEAD}$, [his friend]$_{x\text{-}ATTRIB}$

(52) [His friend]$_{x\text{-}HEAD}$, [the sheriff]$_{x\text{-}ATTRIB}$

For the purpose of determining relative specificity, specific names of diseases and technologies are classified as proper names, whether they are capitalized or not.

(53) [A dangerous bacteria]$_{x\text{-}ATTRIB}$, [bacillium]$_{x\text{-}HEAD}$, is found...

Appositives consisting of more than two NPs include only one head, and multiple attributes.

(54) [Robert V. Van Fossan]$_{X\text{-}HEAD}$, [63]$_{X\text{-}ATTRIB}$, [the chairman of Mutual Benefit Life Insurance Co.]$_{X\text{-}ATTRIB}$

3.2 Linking appositive spans to other referents

Only the single span containing the entire appositive construction is, in turn, eligible to be linked in an IDENT chain. In the example below, the entire span can be linked to later mentions of

"Richard Godown." The two sub-spans, [Richard Godown] and [president of the Industrial Biotechnology Association], are **not** included as separate links in the IDENT chain, since they already form an APPOS chain.

(55) [[Richard Godown]$_{x\text{-}HEAD}$, [president of the company]$_{x\text{-}ATTRIB}$]$_{y}$, gave a speech today. [He]$_{y}$ said...

- APPOS chain X: [Richard Godown]$_{HEAD}$, [president of the company]$_{ATTRIB}$
- IDENT chain Y: [Richard Godown, president of the company], [He]

3.3 Special cases

Adjacent spans containing equivalent amounts of money in different currencies are marked as appositives.

(56) [50 million Canadian dollars]$_{HEAD}$ ([US$ 42.5 million]$_{ATTRIB}$)

Appositives that contain adverbs are marked, **as long as** the adverb does **not** affect the scope or size of the entity. In (57), "the OTC market" and "a base for the small investor" have the same scope, despite the presence of the adverb "traditionally," and can be linked as an appositive construction. In (58), however, the adverb "primarily" narrows the scope from *any* "outside vendors" to *three specific* vendors, eliminating the possibility for co-reference.

(57) The problem has been particularly damaging to [the OTC market]$_{X\text{-}HEAD}$, *traditionally* [a base for the small investor]$_{X\text{-}ATTRIB}$

- APPOS chain: [the OTC market]$_{X\text{-}HEAD}$, [a base for the small investor]$_{X\text{-}ATTRIB}$

(58) Gulf Power had set up an elaborate payment system through which it reimbursed [outside vendors] – *primarily* [three Florida advertising agencies] – for making illegal political contributions on its behalf.

Numeric ages are interpreted as elliptical constructions of a full noun phrase, for example, "42" is an ellipsis of "a 42-year-old." These are marked as appositives.

(59) [Mr. Smith]$_{X\text{-}HEAD}$, [42]$_{X\text{-}ATTRIB}$,

(60) [Three children]$_{X\text{-}HEAD}$, [ages 2, 5, and 10]$_{X\text{-}ATTRIB}$

Job titles are **not** marked as attributes, except when preceded by a definite article or a possessive:

(61) [Secretary of State Colin Powell]

(62) [Yugoslavian President Vojislav Kostunica]

(63) [the Secretary of State]$_{X\text{-}ATTRIB}$ [Colin Powell]$_{X\text{-}HEAD}$

(64) [Yugoslavia's President]$_{X\text{-}ATTRIB}$ [Vojislav Kostunica]$_{X\text{-}HEAD}$

4 Special Issues

4.1 Organization and members

No co-reference is marked between an organization and a subset of its members.

(65) It was an ideal place for [the Orange Workers]$_x$ to start [their]$_x$ new nation, unencumbered by the demographics that have undermined apartheid elsewhere in South Africa. So far, [about 150 Orange Workers] have moved here.

- IDENT chain: [the Orange Workers], [their]
- [about 150 Orange Workers] (subset: no co-ref.)

4.2 Gender and Number

If there is a disagreement in number or gender, yet both noun phrases clearly refer to the same entity, it is acceptable to link a singular NP to a plural, or a masculine NP to a feminine.

(66) And lawmakers are putting the finishing touches on a compromise that would give the Air Force nearly all of the $ 2.4 billion it wants for production of [Northrop Corp.'s radar-eluding B-2 bombers, which cost $ 530 million apiece]$_x$. The final [B-2]$_x$ agreement is certain to require detailed testing and verification of [the bomber's]$_x$ capabilities.

- IDENT chain: [Northrop Corp.'s radar-eluding B-2 *bombers*, which cost $ 530 million apiece](plural), [B-2], [the *bomber's*](singular)

4.3 Indefinite uses of proper nouns

When a proper noun is used as an indefinite reference, it is **not** eligible for co-reference. In (67), "a Hungary" **cannot** be linked to the other indefinite reference, or to the IDENT chain containing the two definite references to the country.

(67) Nor is it [a Hungary], where yesterday the parliament approved constitutional changes meant to help turn [the Communist nation]$_x$ into a multiparty democracy...Erasing the differences still dividing Europe, and the vast international reordering that implies, won't endanger the statehood of [a Poland] or [a Hungary]...With this year's dislocations in China and the Soviet Union, and the drive to democracy in [Poland] and [Hungary]$_x$, the East German leadership grew still more defensive.

- IDENT chain: [the Communist nation], [Hungary]
- [a Hungary] (indefinite: no co-ref.)

4.4 GPEs and governments

GPEs are linked to references to their governments, including metonymic mentions of the capital city, even when the references are nested NPs, or the modifier and head of a single NP.

(68) Christian legislators are insisting on a Syrian troop pullout from [Lebanon]$_x$ before agreeing to political changes giving [the nation's]$_x$ Moslems a greater role in [[Beirut's]$_x$ government].

(69) IDENT([Lebanon], [Beirut's government], [the nation's], [Beirut's])

However, GPEs are **not** linked to mentions of their populations

(70) During Milosevic's 13 years of power, [the people of [Yugoslavia]$_y$]$_x$ saw [[their]$_x$ country]$_y$ torn apart.

4.5 Quantifying Expressions

4.5.1 Quantifiers

Quantifying expressions, sometimes called '*of* expressions,' consist of an entity or group/set of entities (Y) that is modified by a quantifier (X): "X of Y." Quantifiers are words that express some quantity (or 'set'). Examples include cardinal numbers (*five, a billion, hundreds*), partitives (*some, few, half*), measurements (*a gallon, a handful*), and collective nouns (*a herd, a troop*). A quantifying expression should **not** be co-referenced with the entity it modifies.

(71) [three of [them]]$_x$ (cardinal number)

(72) [a lot of [nonsense]]$_x$ (partitive)

(73) [a stretch of [highway]]$_x$ (measurement)

(74) [a flock of [linguists]]$_x$ (collective noun)

In these examples, the larger span is not co-referenced with the smaller span. The larger span can, however, be co-referenced with any subsequent mentions of the same entity. The smaller span is only eligible for co-reference with other mentions under certain circumstances (see section 4.5.3 below).

4.5.2 Partitives

Partitives are quantifiers that refer to a subset (or 'part') of a larger set of entities. With partitives, the larger span is co-referenced with other mentions, but there is no co-reference between the smaller span (the entire set of entities) and the larger span (the subset picked out by the partitive), since these two sets are not equivalent:

(75) [a group of [doctors]]$_x$

(76) [a bunch of [flowers]]$_x$

(77) [a number of [American citizens]]$_x$

(78) [a pinch of [salt]]$_x$

The only exceptions to this rule are words such as "all" and "both," which pick out the entire set of entities, rather than a smaller subset. In this case the two sets of entities **are** equivalent, and **should** be co-referenced:

(79) [All of [the scientists]$_x$]$_x$ spoke at the meeting

- IDENT chain: [All of the scientists], [the scientists]

(80) It was a gift for [both of [us]$_x$]$_x$

- IDENT chain: [both of us], [us]

4.5.3 Linking quantifying expressions to other mentions

The larger span of a quantifying expression is always eligible for co-reference with any subsequent mentions of the same entity:

(81) I ordered [a cup of [coffee]]$_x$ but [it]$_x$ never arrived.

When a quantifier or partitive refers to a subset of a **non-generic** group, the subset and the larger group are both eligible for co-reference with subsequent mentions, though not with each other, resulting in two separate IDENT chains.

(82) [Five doctors]$_x$ presented [their]$_x$ research, and then [three of [**the doctors**]$_x$]$_y$ offered [their]$_y$ opinions

- IDENT chain X: [five doctors], [their], [the doctors](specific group)

- IDENT chain Y: [three of the doctors], [their]

(83) [Half of [**the Palestinian population**]$_x$]$_y$

(84) [Most of [**the attendees**]$_x$]$_y$

If the larger group is **generic** (including all bare plurals) or underspecified, it is **not** eligible for co-reference. The subset, however, should still be linked to any subsequent mentions.

(85) [A group of [**doctors**]]$_x$ offered [their]$_x$ opinions, but [**doctors**] are often known to disagree.

- IDENT chain: [A group of doctors], [their]

- [doctors] (generic: no co-ref.)

(86) [A handful of [**Palestinians**]$_{\text{no coref}}$]$_x$

(87) [A busload of [**attendees**]$_{\text{no coref}}$]$_x$

(88) [dozens of [**friends**]$_{\text{no coref}}$]$_x$

4.6 Possessive extents

Noun phrases extracted from the treebank may include the possessive *'s* in the NP. The *'s* ending should be included in the extent of the noun phrases that are co-referenced.

(89) [Iowa's]$_x$ governor spoke in Postville, [Iowa]$_x$ today.

4.7 Formulaic mentions

In broadcast news, the reporter's introduction and sign off are often structured as follows:

(90) Introduction: [ABC's]$_x$ Jim Sciutto reports from Postville

Sign off: Jim Sciutto , [ABC News]$_x$, Postville, Iowa

In these cases, [ABC's] and [ABC News] are interpreted as the same entity and are co-referenced.

4.8 Sentence fragments

When appositive-like mentions appear in adjacent sentence fragments, these should be annotated as IDENT co-reference.

(91) [The price]$_x$? [$ 300]$_x$. [A lot]$_x$ by current standards.

4.9 Metonyms

Metonymic mentions, in which the name of a location is commonly associated with some larger concept or activity based at that location, can be linked to one another

In (92), "Washington" refers to the U.S. government, and so should be linked to other mentions of the United States as a GPE. In (93) "The White House" refers to the executive branch of the U.S. government, and should be linked to other mentions of the President and his administration (although **not** to "Bush" alone).

(92) [Washington]$_x$ believes North Korea may have enough nuclear fuel to make more than eight or nine atomic weapons...[The U.S.]$_x$ continues to push for disarmament in the region.

- IDENT chain: [Washington](metonym), [The U.S.]

(93) [The Bush administration] was forced to do an about-face last week, after a proposal [it] had made to restrict student-loan borrowers from consolidating their loans at a low, fixed interest rate ran into a firestorm of criticism. [The White House] withdrew the proposal just a few days after offering it.

- IDENT chain: [The Bush administration], [it], [The White House] (metonym)

However, mentions clearly referring to the location and nothing more (usually preceded by a locative preposition, such as *in*, *at*, *to, from*) **cannot** be interpreted as metonyms, and should only

be linked to other locative mentions. In (94), "Washington" cannot be coreferenced with "the country," and "the White House" cannot be coreferenced with "The Bush administration."

(94) [The Bush administration] maintains that the President has [the country's] best interests at heart. This is Kelli Arena in [Washington], reporting live from [the White House].

Similarly, the facility where an organization meets can often be referred to by the name of the organization alone. In this case, the name should be understood as a reference to the facility, **not** the organization, and can thus **only** be linked to other mentions of the facility. Mentions of the organization should be linked in a separate chain.

(95) VOA correspondent Breck Ardery reports from [the [UN]$_y$ headquarters]$_x$. The chief Palestinian observer to [the United Nations]$_y$, Nasir Al-Kidwa spoke briefly today, followed by Israeli [UN]$_y$ Ambassador Yehud Lancry. Breck Ardery , VOA News at [the United Nations]$_x$.

- There are two separate IDENT chains, one (X) referring to a location (the UN headquarters), and the second (Y) to the organization itself.

Supplement A

OVERVIEW

Humans are able to understand a text using knowledge of language as well as knowledge of the world. To get a computer to do the same, it is helpful to prepare examples of text marked up with whatever information the human needed to extract from it. The resulting corpus of annotated examples can then be used to teach the computer to automatically extract the same kind of information.

For this OntoNotes task, the goal is to teach the computer to automatically identify multiple mentions of the same entity or set of entities, including events. In order to do this, annotators will be creating a corpus in which all the instances, or 'mentions,' of each particular entity have been linked. Such groups of linked mentions are called **co-reference chains**. Co-reference chains are made up of noun phrases (NPs), including nominals, pronouns, possessives, proper noun premodifiers (PreMods), and under some circumstances, verbs. These categories will be explained in further detail below.

REFERENCE

Before deciding whether any two mentions are co-referential, annotators must first determine whether the mentions themselves are actually referential.

A linguistic expression is considered referential if a speaker intends it to pick out some particular, **independently distinguishable** entity or set of entities in the world. This entity or set of entities is called the **referent** of the expression. By 'independently distinguishable,' we mean that the entity can be identified using properties other than those inherent in the meaning of the expression itself. Only when the speaker has a specific entity in mind is that entity independently distinguishable from other entities of the same type.

For example, by itself the mention "car" or "a car" is not referential. It could be any vehicle of the type 'car,' and has no inherent properties beyond those common to all cars. However, the mention "my sister's car" is referential. The speaker intends it to refer to one specific car with many properties, such as color, make, and model, distinguishing it from other members of the 'car' category.

Often, the context or form of the linguistic expression determines whether it is referential. By themselves, expressions that are ambiguous, indeterminate, non-specific, generic, or part of a fixed expression do not refer to any particular entity or set of entities in the world, and so cannot be linked together to form a co-reference chain.

(adapted from *The Cambridge Grammar of the English Language* and the Georgetown Guidelines for Protein Name-Tagging, 2004)

TYPES OF MENTIONS

Noun phrases (NPs)

A noun phrase consists of a noun or pronoun, any determiners (*a, an, the, etc.),* and modifiers such as adjectives. All noun phrases are extracted from previously TreeBanked data and

presented to annotators as highlighted spans prior to annotation. Except in specific cases, annotators do **not** need to identify or manually extract any additional spans. Whenever head-sharing NPs are nested, the largest logical span is used in co-reference.

Premodifiers (PreMods)

A premodifier is a word that precedes and modifies a noun. In the noun phrase "the Obama administration," 'Obama' is a proper noun PreMod. Proper noun PreMods can be co-referenced to existing noun phrases **and/or** other proper PreMods, and should be manually extracted by the annotator and added to the list of mentions. Non-proper and adjectival premodifiers are **not** eligible for co-reference.

Verbs

Verbs can be manually extracted by the annotator and added as single-word spans **if** (and only if) they can be co-referenced with an existing noun phrase. Only the single-word head of the verb phrase is included in the span, even in cases where the entire verb phrase is the logical co-referent.

TYPES OF CO-REFERENCE CHAINS

Identity (IDENT)

Names, nominal mentions, pronominal mentions, and verbal mentions of the same entity, concept, or event are co-referenced as IDENT, as their referents all have the same identity. There is no restriction on which semantic types can be considered for co-reference; in particular, co-reference is **not** limited to ACE types.

Appositive (APPOS)

Appositive constructions consist of two (or more) immediately-adjacent noun phrases, separated only by a comma, colon, dash, or parenthesis. The first NP is usually the referential element, called the HEAD, which points to a specific object/concept in the world, and the adjacent NP(s) specify one or more ATTRIBUTES of that referent, thus renaming or further defining the head.

Supplement B

NON-REFERENTIAL MENTIONS

As stated earlier, not all mentions have reference, and non-referential mentions are, by definition, not eligible for co-reference. Below are several categories of such mentions.

Denotation

The denotation of a linguistic expression is its meaning. Dictionaries are concerned with denotation, not linguistic reference. Mentions that are denotational cannot be used in co-reference. Such mentions are not independently distinguishable, as they contain no information beyond the meaning of the expression in the language system.

In the example below, both mentions of [cataract surgery] simply denote a type of surgery, rather than referring to any particular instance of such a procedure, and therefore cannot be linked

> *Allergan Inc. said it received approval to sell the PhacoFlex intraocular lens, the first foldable silicone lens available for [cataract surgery]. The lens' foldability enables it to be inserted in smaller incisions than are now possible for [cataract surgery].*

Negatives

Mentions preceded by a negative cannot be used in co-reference, as they literally refer to an empty or null set. In the example below, no co-reference chain can be created, as [No students] refers to an empty set, and [They] refers to an implicit set that has not been overtly mentioned.

> *[No students] arrived on time. [They] had all overslept.*

However, if the expression involves a quantifier, the set evoked can then be used in co-reference. In the example below, the set [the students] has been explicitly mentioned and can thus be co-referenced with [They]. [None of the students] remains unlinked.

> *[None of [the students]$_X$] arrived on time. [They]$_X$ had all overslept.*

Similarly, mentions that are governed by negated verbs cannot be used in co-reference:

> *My sister doesn't have [a car].*

Determinatives

Determiners such as "each" and "either" evoke an implicit set of entities, without actually referring to that set. In the example below, [Either city] cannot be co-referential with [They], as the first mention can refer to any one, *but not both*, of the cities, while the second *must* refer to the implicit set of both cities.

> *[Either city] might win the Olympics. [They] have excellent facilities.*

Again, however, if the expression involves a quantifier, co-reference becomes possible.

[Either of [the cities]$_X$] might win the Olympics. [They]$_X$ have excellent facilities.

POTENTIALLY NON-REFERENTIAL MENTIONS

Context often plays a role in determining whether a mention is referential or not. The same words may be referential in one context, but non-referential in another. Often, a later mention will clear up the ambiguity.

Indeterminate Mentions

Mentions are considered indeterminate if it is impossible, without any further context, to tell whether the speaker has a particular referent in mind. In the two examples below, it may be the case that the speakers have no idea who wrote the email or stole the CD player:

[The boy who wrote this email] must be expelled.
I think Ed's CD player was stolen by [a friend of his].

On the other hand, the speakers in the examples above may have had a particular boy and a particular friend in mind when they spoke. Because the later sentences provide a broader context for the mentions, co-reference becomes possible.

[The boy who wrote this email]$_x$ must be expelled. I will write to [his]$_x$ parents immediately.

I think Ed's CD player was stolen by [a friend of his]$_x$. [He]$_x$ is the only other person with a key to Ed's apartment.

Non-Specific Mentions

Indefinite noun phrases, which begin with the indefinite article (*a, an*), are often non-specific. In the first example below, it is impossible to tell without any further context whether the speaker has a specific Norwegian in mind, and so the mention should be considered non-specific.

I intend to date[a Norwegian].

Indefinite noun phrases become even more ambiguous when governed by a determinative, as in the example below, which, depending on the context, might not refer to any particular film star, or might refer to one specific film star, or one film star for each of the people in question:

Each of them wants to marry [a film star].

However, as with Indeterminate Mentions, if the later context makes it obvious that the speaker does have a specific Norwegian or film star(s) in mind, later mentions can be coreferenced:

I intend to date [a Norwegian]$_x$. I met [him]$_x$ at a dinner party last week.

Each of them wants to marry [a film star]$_X$, and [his]$_X$ name is Brad Pitt.

Generic Mentions

Generic mentions are those that refer to a class, type, or kind of entity, rather than to any specific members of that class. Often, they occur as "bare plurals," which do not contain an article.

In the example below, the first mention of "lions" cannot be linked to the second mention, because both refer to lions as a type of creature, rather to than any, or even all, specific members of that species.

> *[Lions] are ferocious beasts. However, [lions] do not usually hunt people.*

The example below makes it clear that the first mentions of "young black men" and "Hispanic kids" cannot possibly be co-referent with the second ones, although both are mentions of the same generic set of people.

> *One of the major gang problems -- of the insanity of gang problems is [young black men] killing [young black men] and [Hispanic kids] killing [Hispanic kids].*

However, generic mentions ARE co-referent with any subsequent pronouns. In the example below, [Lions] and [they] should be linked (X), but the second generic mention of [lions] should NOT be included in the same co-reference chain. Rather, a new co-reference chain (Y) must be created for [lions] and [Their].

> *$[Lions]_X$ are ferocious beasts. However, $[they]_X$ do not usually hunt people. Instead, $[lions]_Y$ usually prey on medium to large-sized mammals. $[Their]_Y$ main prey includes antelope and zebra.*

ARTICLES AS A TEST FOR REFERENTIALITY

<u>The Definite Article</u>

Mentions that begin with the definite article (*the*), are almost always referential. A few exceptions are listed below:

Class Uses: The definite article can be used to refer generically to a species of animal.

> *[The African elephant] will soon be extinct*

Fixed Expressions: The definite article occurs in some non-referential expressions concerning musical instruments, diseases, transportation, etc.

> *Wolfgang can play [the piano]*
> *I have [the flu]*
> *We took [the train]*
> *I spoke to her on [the telephone]*
> *He was in [the hospital] for a week*

Job Positions: Any mention of the job position itself is non-referential, while a reference to the current occupant of the position is referential. Even though both examples below use the definite article, the first is non-referential, while the second is referential.

> *[The president of the United States] has been assassinated three times.*
> *[The President of the United States] just arrived on Air Force One.*

The Indefinite Article

Most mentions that begin with indefinite articles (*a, an*) are at least potentially, if not certainly, non-referential. Some examples not discussed above include:

Expressions of Rate and Measurement: Any unit of measurement (time, length, etc.), including those used in expressions of rate, is considered non-referential.

> *She has a salary of $80,000 [a year]*
> *Her house is [a mile] from here.*

Indefinite Proper Nouns: When a proper noun is preceded by an indefinite article, it is no longer considered referential. In the example below, the meaning of "a Hungary" is "a country *like* Hungary," rather than a reference to Hungary itself.

> *Erasing the differences still dividing Europe, and the vast international reordering that implies, won't endanger the statehood of [a Poland] or [a Hungary]…*
>
> *With this year's dislocations in China and the Soviet Union, and the drive to democracy in [Poland] and $[Hungary]_x$, the East German leadership grew still more defensive.*

One exception to the rules of indefinite mentions occurs when a news anchor offers a brief overview, similar to a headline, before discussing the details of a story. This often results in two sequential indefinite mentions of the entities involved: once in the introduction and once in the story itself. In this instance alone, the entities in the introduction **should** be linked to the corresponding mentions within the story:

> *"Next, $[a\ man]_x$ robs $[a\ bank]_y$ at gunpoint, but leaves $[[his]_x\ wallet]_z$ behind. Late yesterday afternoon, $[a\ man]_x$ posing as a customer demanded money from the teller at $[a\ local\ Bank\ of\ America\ branch]_y$. $[The\ suspect]_x$ was later apprehended at home by police, who found $[his]_x$ name and address in $[a\ wallet]_z$ recovered at the scene."*

NPs with No Article

So-called "bare NPs" are mentions that do not begin with an article or a possessive. They are almost always either potentially or certainly non-referential.

Bare Plurals: These are typically generic mentions

> *$[Officials]_x$ said $[they]_x$ are tired of making the same statements.*

[Meetings]$_x$ are most productive when [they]$_x$ are held in the morning. [Those meetings]$_x$, however, generally have the worst attendance.

[Parents]$_x$ should be involved with their children's education at home, not in school. [They]$_x$ should see to it that [their]$_x$ kids don't play truant; [they]$_x$ should make certain that the children spend enough time doing homework; [they]$_x$ should scrutinize the report card. [Parents]$_y$ are too likely to blame schools for the educational limitations of [their]$_y$ children. If [parents]$_z$ are dissatisfied with a school, [they]$_z$ should have the option of switching to another.

Fixed Expressions or Frames: These are often locatives, describing what *type* of location a person is at, without referring to a particular building, facility, etc. In the examples below, mentions of "high school" and "camp" should not be co-referenced with the actual names of the school or the camp.

Do you know [George Washington High]? I went to [high school] there.
Bob is at [camp] for the summer. He always enjoys [Camp Greentree].

NON-REFERENTIAL PRONOUNS

Generic *you*, *we*, *they*

Instances of the generic you, your, and yourself are not linked. In the following example, none of the four occurrences of the pronoun "you" are eligible for coreference:

A lot of times, Frist recalls, [you]'d have a critical patient lying there waiting for a new heart, and [you]'d want to cut, but [you] couldn't start unless [you] knew that the replacement heart would make it to the operating room.

Expletive Pronouns

Pronouns used as expletives (sometimes called 'dummy' pronouns) are also not linked. These pronouns do not refer to anything, and so cannot be linked to another noun phrase. In the example below, 'it' and 'there' are expletive pronouns, and should remain unlinked.

Since [there] have been no further negotiations, [it] seems obvious that the violence will continue.

As a test for expletive pronouns, remember that normal pronouns can be replaced with a noun phrase, while expletive pronouns cannot.

Supplement C
TESTS FOR GENERIC-HOOD

The distinction between a generic mention and a specific one is often very subtle. There are no hard-and-fast rules about generic-hood, but there are several tests annotators can do on a particular example in order to determine whether it likely to be generic.

1. **The Determiner Test**
   - A singular noun with an indefinite article (a/an) as its determiner is likely to be generic.
     - "I think I'll buy [a dog]."
       - The speaker doesn't have a particular dog in mind, any dog might be the one she ends up buying.
   - A plural noun with NO determiner (a 'bare plural') is likely to be generic.
     - "I like [dogs]."
       - The speaker is making a generalization about all dogs, not any particular dog or group of dogs.

2. **The Pronoun Test**
   - If the noun in question can be replaced with a pronoun, it is likely to be specific, NOT generic.
     - "I just talked to Bob on [the phone]."
     - Pronoun replacement: "I just talked to Bob on ***it***."
       - The utterance is strange if the intended meaning is "I just called Bob via telephone and we talked"
     - But, the speaker may have been talking about a particular phone: "I just talked to Bob on $[\text{the phone}]_X$, but now $[\text{it}]_X$ seems to be broken."
       - The pronoun replacement sounds fine because the noun is not generic

3. **The Paraphrase Test**
   - If the noun in question can be paraphrased using a different part of speech (an adverb, verb, adjective, etc.), it is likely to be generic.
     - "She was in [the hospital] for three weeks."
     - Paraphrase: "She was ***hospitalized*** for three weeks."
       - There is a good chance this mention is generic.
     - The speaker may have had a specific hospital in mind: "She was in $[\text{the hospital}]_X$ for three weeks – $[\text{it}]_X$ was Mass General."
       - Now the paraphrase sounds strange because the noun is not generic: "She was ***hospitalized*** for three weeks – $[\text{it}]_?$ was Mass General."

4. **The Negation Test**
   - Negated nouns and pronouns are generic – they literally point to a referent that does not exist.
     - "I saw [**no** people] in the room"
       - Compare with the specific mention: "I saw [some people]$_X$ in the room , and [they]$_X$ were dancing.
     - "I saw [**no** one] in the room"
       - Compare with the specific mention: "I saw [one person]$_X$ in the room, and [he]$_X$ was dancing."
     - "I do **not** have [a dog]"
       - Compare with the specific mention: "I have [a dog]$_X$ and [his]$_X$ name is Fido.

5. **The "Boilerplate" or "Empty Shell" Test**
   - "Empty Shell" nouns are usually job positions. Any reference to the job position itself is usually generic, while a reference to the current occupant of the position would be specific.
     - According to the Constitution, [The President of the United States] has the power to veto bills sent to him by Congress.
       - This could refer to any person holding the office of President, and is thus generic
     - [The President of the United States] just arrived in Air Force One.
       - This refers specifically to the person who is President at the time the sentence is uttered

****Remember****

Just because a noun is generic does not mean it will NEVER be linked to anything else: speakers can always refer back to certain generic mentions with pronouns or other types of specific mentions.

- I like [dogs]$_X$ because [they]$_X$ are playful and loyal.
- [No one]$_X$ should have to lose [their]$_X$ home.

However, there are almost NO instances in which we should link one generic mention to another.

- I like [dogs]$_X$. [Dogs]$_Y$, however, do not seem to like me.

Supplement D

COPULAR STRUCTURES

A copular structure consists of a referent (usually the subject), an attribute of that referent (usually the predicate), and a copula that serves to equate (or link) the referent with the attribute. In the example below, [John] is the referent, [a linguist] is the attribute, and "is" is the copula.

> *[John]* ***is*** *[a linguist]*

Annotators should not mark the relationships signaled by copular structures, as they are not co-referential in nature, and will be captured later through word sense tagging.

Recognizing Copulae

Almost all copular structures contain a 'linking' verb, such as *be, appear, feel, look, seem, remain, stay, become, end up, get, etc*. In the following example "called" is the copula.

> ***Called*** *[Otto's Original Oat Bran Beer], [the brew] costs about $12.75 a case.*

However, some copular structures do not include a verb. In the first example below, "or" functions as a copula. In the second example, the construction is interpreted as having an implicit copula.

> *The sale will facilitate Gen-Probe's marketing of a diagnostic test for [acquired immune deficiency syndrome],* ***or*** *[AIDS].*
>
> *John considers [Fred] [an idiot]. (=John considers Fred* ***to be*** *an idiot.)*

Identifying the Referent

Only the most specific element of a copular structure, the referent, should be linked to any subsequent mentions. This is most often, though not always, the subject. The attribute, usually the predicate, remains unlinked. When in doubt, annotators should refer to the same specificity scale used for determining the heads of appositives. If both elements are of equal specificity, the leftmost element should be used for co-reference.

> **(MOST SPECIFIC)----------(LEAST SPECIFIC)**
> proper noun **>** pronoun **>** definite NP **>** indefinite specific NP **>** non-specific NP
> John **>** he **>** the linguist **>** a linguist **>** noted linguist

Based on this scale, in both examples below [John Smith], as a proper noun, is the referent of the copular structure and should be linked to subsequent mentions of [him], [he] and [his].

> *$[John\ Smith]_x$ is a [linguist]. People are nervous around $[him]_x$, because $[he]_x$ always corrects their grammar.*

[The president of the bank] is [John Smith]$_X$. [He]$_x$ loves [his]$_x$ job.

Supplement E

APPOSITIVES

An appositive construction contains a noun phrase that is modified by one or more immediately-adjacent noun phrases. The elements are generally separated by only a comma, colon, or parenthesis.

[John]$_{x\text{-}HEAD}$,[a linguist I know]$_{x\text{-}ATTRIB}$, is coming to dinner.

Identifying the Head and Attribute(s)

APPOS chains consist of a HEAD, which is referential, and one or more ATTRIBUTES, which rename or further describe the HEAD. For each appositive construction, the most specific element is marked as the HEAD.

(MOST SPECIFIC)----------(LEAST SPECIFIC)
proper noun **>** pronoun > definite NP **>** indefinite specific NP **>** non-specific NP
John **>** he **>** the linguist **>** a linguist I know **>** noted linguist

In the examples below, the underlined element is the most specific, and therefore is marked as the head of the construction

[<u>John Smith</u>]$_{x\text{-}HEAD}$, [noted linguist]$_{x\text{-}ATTRIB}$,

[A famous linguist]$_{x\text{-}ATTRIB}$, [<u>he</u>]$_{x\text{-}HEAD}$ studied at MIT

[the president of the linguistics club]$_{x\text{-}ATTRIB}$, [<u>J. Smith</u>]$_{x\text{-}HEAD}$

In cases where the two members of the APPOS chain are equivalent in specificity, as in the three examples below, the leftmost member of the appositive is marked as the HEAD.

[<u>The chairman</u>]$_{x\text{-}HEAD}$, [the man who never gives up]$_{x\text{-}ATTRIB}$

[<u>The sheriff</u>]$_{x\text{-}HEAD}$, [his friend]$_{x\text{-}ATTRIB}$

[<u>His friend</u>]$_{x\text{-}HEAD}$, [the sheriff]$_{x\text{-}ATTRIB}$

Appositives consisting of more than two NPs include only one head, and multiple attributes.

[<u>Robert V. Van Fossan</u>]$_{X\text{-}HEAD}$, [63]$_{X\text{-}ATTRIB}$, [the chairman of Mutual Benefit Life Insurance Co.]$_{X\text{-}ATTRIB}$

Linking appositive spans to other referents

Only the single span containing the entire appositive construction is, in turn, eligible to be linked in an IDENT chain. In the example below, the entire span can be linked to later mentions of "Richard Godown." The two sub-spans, [Richard Godown] and [president of the Industrial Biotechnology Association], are **not** included as separate links in the IDENT chain (Y), since they already form an APPOS chain (X).

[[Richard Godown]$_{x\text{-}HEAD}$, [president of the company]$_{x\text{-}ATTRIB}$]$_{y}$, gave a speech today. [He]$_{y}$ said...

SPECIAL CASES FOR APPOSITIVES

Names of diseases and technologies are considered to be as specific as proper names, whether they are capitalized or not.

[A dangerous bacteria]$_{x\text{-}ATTRIB}$, [bacillium]$_{x\text{-}HEAD}$, is found...

Equivalent amounts of money in different currencies contained in adjacent spans are marked as appositives, as are equivalent times in different time zones.

[50 million Canadian dollars]$_{HEAD}$ ([US$ 42.5 million]$_{ATTRIB}$)

Appositives that contain adverbs are marked, as long as the adverb does NOT affect the scope or size of the entity. In the first example below, the adverb "traditionally" does not affect the size of the entity it modifies, that is, [the OTC market] and [a base for the small investor] have the same scope, and so the appositive can be marked.

*The problem has been particularly damaging to [the OTC market]$_{X\text{-}HEAD}$, **traditionally** [a base for the small investor]$_{X\text{-}ATTRIB}$*

However, in the second example below, the adverb "primarily" does affect the size of the entity it modifies: [outside vendors] is larger in scope than [three Florida advertising agencies], and so the appositive cannot be marked.

*Gulf Power had set up an elaborate payment system through which it reimbursed [outside vendors] - **primarily** [three Florida advertising agencies] - for making illegal political contributions on its behalf.*

Numeric ages are interpreted as elliptical, or abbreviated, constructions of a full noun phrase, for example, "42" is an ellipsis of "a 42-year-old." These are marked as appositives.

[Mr. Smith]$_{X\text{-}HEAD}$, [42]$_{X\text{-}ATTRIB}$,

[Three children]$_{X\text{-}HEAD}$, [ages 2, 5, and 10]$_{X\text{-}ATTRIB}$

Job titles are NOT marked as attributes, except when preceded by a definite article or a possessive:

[Secretary of State Colin Powell]

[Yugoslavian President Vojislav Kostunica]

[the Secretary of State]$_{X\text{-}ATTRIB}$ [Colin Powell]$_{X\text{-}HEAD}$

[Yugoslavia's President]$_{X\text{-}ATTRIB}$ [Vojislav Kostunica]$_{X\text{-}HEAD}$