Computer Science and Artificial Intelligence Laboratory
Technical Report

MIT-CSAIL-TR-2010-025 May 12, 2010

Annotation Guide for the UCM/MIT Indications, Referential Expressions, and Coreference Corpus (UMIREC Corpus)

Mark Alan Finlayson and Raquel Herv^¡s

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

CSAIL

Annotation Guide for the UCM/MIT Indications, Referential Expressions, and Coreference Corpus

(UMIREC Corpus)

Version 1.0

Mark Alan Finlayson
Computer Science and
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA, 02139 USA
markaf@mit.edu

Raquel Hervás
Departamento de Ingenieria
del Software e Inteligencía Artificial
Universidad Complutense de Madrid
Madrid, 28040 Spain
raquelhb@fdi.ucm.es

Introduction

This is the annotation guide given to the annotators who created the UCM/MIT Indications, Referring Expressions, and Coreference (UMIREC) Corpus version 1.0. The corpus comprises texts annotated for referring expressions, coreference relations between the referring expressions, and so-called "indication structures", which split referring expressions into constituents (nuclei and modifiers) and mark each constituent as either 'distinctive' or 'descriptive', which indicate whether or not the constituent contains information required for uniquely identifying the referent.

The contents of this corpus, the annotation procedure, and the indication structures are described in more detail in a paper titled "The Prevalence of Descriptive Referring Expressions in News and Narrative" published in the proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, held in July 2010 in Uppsala, Sweden (ACL-2010).

Referential Expressions & Coreference

What is a Referent?

Referents are objects that are *referred to* in the text. Most of the time, referents are *things that exist*. Consider the following simple sentence, where the referents have been underlined:

(1) John kissed Mary.

In this sentence, both referents are people – concrete things in the world. While this simple example covers a large number of cases, they are far from the full story. First off, referents may or may not have physical existence:

(2) John had an idea.

Here, the second referent is an abstract object. Similarly, we can refer to things that don't exist:

(3) If John had a car$_1$*, it*$_1$ *would be red.*

The car does not exist, and yet we still refer to it. This sentence also illustrates an important point, namely, that a single referent can be mentioned several times in a text. In (3), it is the car that is mentioned twice. In this case, we say there is a single *referent* (the car), with two *referential expressions* (the phrases "a car" and "it"). These are called *coreferential expressions* because they refer to the same referent. In this annotation, we will be marking both referring expressions and the set of referring expression that refer to the same things. As done in this example, we will use numeral subscripts to indicate that the two references refer to the same referent. Determining whether a particular reference refers to a previously mentioned referent, rather than introducing a new referent, will be dealt with in more detail later.

An initial definition for referents is that *referents are things that have been picked out for special attention*. The trick, then, is to determine what 'special attention' means. In some sense, everything mentioned in a text, be it an object, an event, a time, a place, or something else, has been picked out for special attention, because it's being talked about rather than something else of the same kind from the set of all possible things in the world. In (1), we might have marked "kissed" as a referent, since it is something happening, and we chose to talk about that rather than something else. But if everything mentioned is a referent, nearly everything in a text would be marked as a referring expression, which would be almost as uninformative as having nothing marked. What we are really interested in is *reification*, or, roughly speaking, items that find themselves the subjects or objects of verbs. To be more precise, if something is referred to using a *noun phrase*, it should be marked as a referent. Thus we have rule #1: **Mark noun phrases as referring expressions.**

This definition has the convenient property of having us to mark events (such as "kissed" above) only when they are picked out further beyond their use as a verb. Consider the sentence:

(4) John drove Mary to work.

In (4) we have three noun phrases, and we do not mark the driving event, in accordance with our intuition. But if we appended a second sentence:

(5) John drove$_1$ Mary to work. It$_1$ took forever.

We are picking out the act of driving as something interesting to talk about above and beyond its mere mention in the story, and in so doing we used a noun phrase "it" to refer to the event of driving. This forces us to "retroactively" (so to speak) mark referring expressions to the event. Thus rule #2: **Mark referential expressions that corefer with marked referential expressions, even if they would not normally be marked.**

It is important to keep in mind that not all nouns, noun phrases, or noun-substitutes are referential:

(6) You can make ***it*** *in Hollywood.*

Note that the pronoun "it" is non-referential; therefore, we do not mark it. Noun phrases also might not refer to any object *in particular.* Take the following sentence:

(7) Lions are fierce.

Here we are not referring to a *particular* lion, but rather to a class, the set of all lions. These should be marked as references, but are different from particular lions:

(8) Lions$_1$ are fierce. But Leo the Lion$_2$ was the fiercest of all$_1$.

This indicates why it is important to mark generics. In (8), we would like to indicate what Leo the Lion was the fiercest of – namely, "all Lions."

Despite this, we will not mark all generics, since almost everything is described as a member of some class of objects, e.g.,:

(9) Leo$_1$ was a Lion.

Thus generics (the "a Lion" above), like events, should be marked only when they are directly referred to – in other words, when the author intends to pick out the class itself, rather than merely indicating an object is a member of that class. Thus rule #3: **Mark generics as referring expressions only when they are referred to directly.**

Also make sure to mark numeric noun phrases as referring expressions. In (10) there are three noun phrases containing numerals.

(10) In the 1950s the city had 50,000 inhabitants. In 30 years, the population doubled.

Also mark referential pronouns as referring expressions, including possessive and reflexive pronouns.

(11) John$_1$ was a doctor. He$_1$ paid for his$_1$ studies$_2$ by himself$_1$.

In (11), different types of pronouns (possessive, reflexive) correspond to the same referent "John" and must be marked as coreferential. Note that the phrase "his studies" contains two referring expressions: One to John's studying ("his studies) and another to John himself ("his"). Thus rule #4: **Numeric noun phrases and pronouns should usually be marked as referring expressions**.

Referential Extent: Modifiers

It is important to include in a referring expression not only the core noun (or noun phrase) that is doing the referring, but also to include any modifiers to the referring expression. This is so because modifiers can substantially change the nature of the object being referred to. Compare the two following sentences:

(12) Every morning John woke early.

(13) That morning John woke early.

In (12), the noun phrase "every morning" refers to the set of all mornings, but in (13) we are referring to a specific, single morning with the phrase "that morning." Similarly, you should include determiners (14), pronouns (15), adjectives (16), appositives (17), prepositional phrases (18), and other modifiers as part of the referring expression, as in the following examples.

(14) The car was expensive.

(15) His car was expensive.

(16) The red car is expensive.

(17) The car, red as blood, was expensive.

(18) The car in the garage is expensive.

Thus rule #5: **Mark quantifiers , determiners, pronouns, adjectives, adjectival phrases, and other modifiers as part of the referring expression**.

Take into account that modifiers can themselves contain referring expressions. In example (19), where "Kent cigarette" is a modifier of the whole referring expression "Kent cigarette filters", but at the same time a referring expression itself:

(19) Kent cigarette$_1$ filters$_2$ contained asbestos.

Therefore, in example (19) you will mark two referring expression: "Kent cigarette" and "Kent cigarette filters." Sometimes these rules lead you to mark rather large portions of text as referring expressions, with multiple nesting expressions (below, only the largest has been underlined):

(20) Takuma Yamamoto, vice president of Fujitsu Motor's widgets and cogs division since June 1993, was fired yesterday.

The underlined referring expression in (20) contains three internal referring expressions, namely, the car company, the car company's division, and a date.

Determining Coreference Relationships

The rules elaborated above cover what to mark as a referring expression. Once you have determined that some set of tokens is a referring expression that should be marked, your second task to is to determine if it corefers with a previously-introduced referring expression. We have already seen some unambiguous cases of coreference. Let us consider more subtle cases.

Quantification

The first case is that of quantification:

(21) *Every day John woke early. One day he overslept.*

Are the two marked referring expressions here referring to the same day? The answer is no, as the phrase "every day" refers to a set of days (a fairly large set, in fact), and "one day" refers to a particular day. No problems here. But what about:

(22) *Every day the goose laid a golden egg. The woman could hardly wait for the egg.*

Are they the same egg? This is a bit trickier. It's clear that there is more than one egg – in fact, one egg for every day. And it's clear that the woman could hardly wait for each of them. But does "the golden egg" refer to the set of all the eggs? One technique for determining coreference is to vary the quantification of the second referring expression and see if it changes the meaning:

(23) *Every day the goose laid a golden egg.* ***One day****, the woman could hardly wait for the egg.*

In (23) it is clear the second referring expression is to a particular egg and is not coreferential with the first, since the phrase "one day" breaks us out of talking about the things that happened "every day." This indicates the proper way to look at (22) : the phrase "every day" introduces a special context in which an object (the golden egg) is introduced and referenced. The context, in this case, does not continue into the next sentence, so in (22) we conclude that the two referring expressions do **not** corefer. (Note that this context effect is much like in (3) above, where we introduce an imaginary car in an alternate possible world.) This leads us to rule #6: **With quantified references, use variation of quantifiers to test coreference.**

Plural Referring Expressions

Plural references can present some special problems for coreference. Consider these cases:

(24) *The three sons*$_1$ *stared at one another*$_1$.

(25) *Each of the sons*$_1$ *was strong but lazy*.

Although both "at one another" and "each of the sons" are referring to each singular son, at the same time they are referring to all of them. So both references should be considered as coreferences of "the three sons". Thus remember that some quantifiers can produce plural references even though they are referring to a set of singular referents at the same time.

Synecdoche and Metonymy

Another common case is the use of synecdoche or metonymy, figures of speech in which a part of an object, or a closely related object, is used to refer to the whole object:

(26) *The White House*$_1$ *announced a new economic stimulus plan today. The president and his staff*$_1$ *argued that previous efforts had fallen short.*

In this case "The White House" is a closely-related object that is used to stand in for "The president and his staff." Contrast, however:

(27) The owner of the orchard$_1$ often could be found pruning the old trees$_2$ and propping up the young ones$_3$.

In this case, "the old trees" and "the young trees" are not the same as "the orchard" – they are a part of the orchard, but not the same as it. The easiest way to discover this is to substitute one for the other, and determining if the sentence is (a) still well formed, and (b) the meaning remains unaltered. Thus rule #7: **Use the substitution test to determine appropriate coreference relations.**

"Of" Prepositional Phrases

Another class of referring expressions that can be tricky for determining coreference are those of the form "X of Y", e.g.:

(28) This has caused problems among a group of workers.

Does the phrase "a group of workers" contain one referring expression or two (one to the group of workers, and another to the set of all workers)? Consider these similar examples:

(29) Smoking has caused a high percentage of cancer deaths.

(30) Smoking has caused most cancer deaths.

One way of testing this is to try substituting the 'Y' for the 'X', and seeing if the fundamental class of the referent changes. If the class does not change, we have only a single reference. For example, in (28), we substitute "workers" for "group of workers", we will still be talking about people. Thus we have only a single referring expressions. In (29) the overall referring expression is to a percentage, but the internal object of the "of" prepositional phrases are "cancer deaths." These are clearly different fundamental kinds of objects, and so there are two different referring expression. By contrast, in (30), "most cancer deaths" is the same basic type as "cancer deaths", and so we have only a single reference again.

Copular Expressions

Determining coreference can be tricky in copular ("X is a Y") expressions:

(31) John$_1$ was a scientist$_1$.

(32) John$_1$ was the scientist$_1$.

(33) John$_1$ was not a scientist$_2$.

In (31) we know from the very syntax of the sentence that we are describing John as a scientist, and so the second referring expression is to John. In (32), we are describing John as a particular scientist (one perhaps we talked about earlier in the text), and so it is also coreferential. However, the introduction of "not" in (33) breaks the coreferentiality of the sentence, and we have references to two different things.

Indications

What is an Indication?

Indications are additional information attached to the referring expressions that help us to determine the goal of the author when using that particular choice of words to refer to the referent in question.

The extra information annotated for each referring expression are its (1) constituents, (2) each constituent's function, and (3) whether or not it is a copular predicate. The *constituents* of a referring expression are its nuclei and modifiers. Annotating the constituents will involve marking the boundaries of these parts of each referring expression. The *functions* of the constituents, for the purposes of this annotation scheme, can be either **Distinctive**, **Descriptive** or **Other**. Annotating this will involve choosing one of these three tags to apply to each of a referring expression's nuclei and modifiers.

Annotating nuclei

The nucleus (or nuclei) of a referring expression is the set of words that perform the "core" referential function of the referring expression. Consider the following, where the referring expression has been underlined and the nucleus surrounded by square brackets:

(1) The old [king] was wise.

In (1), the noun phrase "the old king" refers to a king. The core of the referential action of the phrase is the word "king," and so that is our nucleus. Syntactically, we could have very well omitted "old" and said "The king was wise." The nucleus is very similar to the idea of the *head* from linguistics: the *head* of a phrase is a portion of a phrase that plays the same grammatical role (relative to the rest of the sentence) as that of the whole phrase. Since most of our referring expressions are noun phrases, the nuclei of our referring expressions will almost always be nouns (or noun phrases) themselves. Rule #1: **Mark the core referential words of each referring expression as the nucleus (usually nouns, and usually the heads of the phrase).**

Every referring expression **must** have at least one nucleus. This means that if a referring expression consists of only a single token, the nucleus must be the whole referring expression:

(2) [He] drove.
(3) [John] slept.

Most referring expressions will have only a single nucleus. In fact, we have yet to find a referring expression that has more than one nucleus. Thus rule #2: **Most referring expressions have one and only one nucleus.**

We have noted that most of the time, a nucleus will be a noun or noun phrase. Keep in mind, though, that referring expressions often include other words that are nouns, but are not the nucleus of the referring expression. Consider:

(4) The [roof] of the house collapsed.

In (4), both “roof’ and “house” are nouns, but “house” is not the nucleus because it is part of a possessive modifier that is providing extra information about the roof – we could have said “The roof collapsed” without significantly changing the meaning, but not “The house collapsed.” The key consideration here is that the nucleus should have the same basic type as the referring expression as a whole. In (4), we are referring to a roof, and therefore we cannot choose “house” as the nucleus of the expression – this changes the basic type of the object referred to from the roof (a part of the house), to the house itself. Thus rule #3: **The nucleus should have the same basic type as the referring expression itself.**

There will sometimes be cases where the nucleus of the referring expression doesn’t look like a normal noun. Consider:

(5) Don’t let the [perfect] be the enemy of the [good].
(6) The [reading] of the will was a sad moment.

In (5) both underlined referring expressions have adjectives (used as nouns) as their nuclei, and in (6) the nucleus “reading” is a verb acting like a noun.

You may have noticed that determiners and other function words have been left out of the nucleus. In (5) we marked “perfect” instead of “the perfect” and in (6) “reading” instead of “the reading”. In marking nuclei, we want to pare down the word selection to the bare minimum core. Thus rule #4: **Exclude as many words as possible from the nucleus.**

Annotating modifiers

Modifiers are those portions of the referring expression that augment or change slightly the meaning of the nuclei. For the most part, anything that is not a nucleus will be a modifier. In the following examples, the modifiers of the referring expressions have been surrounded by parentheses:

(7) (The) (old) [woman] sat and knitted.
(8) (The) (cold) [water] ran clear and bright.

Modifiers may not only be adjectives, but also other kinds of words that provide information about the nuclei. Consider the following (only a single referring expression in each sentence has been underlined):

(9) (The) [roof] (of the house) collapsed
(10) (Her) [mother] wanted the best for her.
(11) (Three) [ghosts] visited him.
(12) He tended the orchard (every) [morning].

In these examples, we have a determiner (9), a prepositional phrase (9), a possessive pronoun (10), a numeral (11), and a quantifier (12) all acting as modifiers. Also note that, just like in (9), a referring expression may have multiple modifiers. Consider (13), where the underlined referring expression has four modifiers:

(13) (The) (three), (strong), but (lazy) [sons] did nothing.

Almost all words that are not in a nucleus are parts of modifiers – there are only a few isolated cases where there is a word in a referring expression that is not in either a nucleus or a modifier, for example, the "but" in (13). Thus rule #5: **Most non-nucleus words, no matter their parts of speech, will be part of some modifier**.

Annotating functions

The function of a nucleus or modifier can be either ***Distinctive***, ***Descriptive***, or ***Other***. These three categories are intended to reflect the goal of the author when constructing the referring expression in question. The *distinctive* function is the default case: it represents the use of a nucleus or modifier to identify a referent. We will assume that all parts of referring expressions are at the very least distinctive. The *other* function is a catch-all category for unclassifiable cases – a nucleus or a modifier will be *other* only when it is neither distinctive nor descriptive. We have not encountered many of these cases so far, but this tag is available to you if you need it. Your main goal, then, is to determine if the constituents of a referring expression are *descriptive*. This criterion is laid out in rule #6: **Mark a nucleus or modifier as descriptive if it contains more information than needed to unambiguously identify the referent of the referring expression.**

To understand what it means to unambiguously identify the referent of a referring expression, let us first consider some ambiguous cases. Consider the following short story:

(14) *John owned a black dog, and Mary a white one. The dog was a barker.*

To which dog does "the dog" refer, John's or Mary's? There is not enough information to be sure. This is an example of an ambiguous referring expression. In contrast, if we were to say:

(15) *John owned a black dog, and Mary a white one. The **black** dog was a barker.*

It is now clear to which dog we are referring. We might also have said, in place of "the black dog," something like "John's dog," "the first dog," "the former," or some other equivalent phrase. These are all examples of referring expressions that give you enough information to identify the referent. In our terms, the nucleus and modifier of this referring expression (as well as the referring expression itself) are distinctive.

Authors often introduce new information inside a referring expression. Consider the following excerpt from a newspaper article (where only the referring expression of interest has been underlined):

(16) *A young student was today fighting for her life after fire ripped through her Edinburgh flat. [Nicola Graham] is in a "serious but stable" condition at the specialist burns unit in St. John's Hospital.*

In (16), the author could have just as well used "She" in place of "Nicola Graham," but the author decided instead to give us the student's name. Thus this referring expression is called a descriptive referring expression.

Because sometimes the new information is given in a modifier, and sometimes in a nucleus, you will be marking each constituent individual either descriptive or distinctive. Consider the following example, wherein constituents with a ‘C’ superscript are descriptive, and constituents with a ‘T’ are distinctive:

(17) *John liked (the)^{T} (white)^{C} [dog]^{T}. Mary liked it too.*

In (17) we don’t really need to know that the dog is white, as there are no other dogs mentioned; thus the referring expression as a whole is descriptive. Because the extra information is given in the modifier, we mark the modifier “white” as descriptive. There is nothing fancy or unusual about referring to a dog as a “dog”, and so the nucleus in (17) is distinctive.

Modifiers provide information about the nucleus of a referring expression. In order to determine if they are *distinctive* or *descriptive* you must determine if the referent would be unambiguous if the modifier is removed. Some examples:

(18) *John preferred (the)^{T} (old)^{T} [car]$^{T}_{1}$, but Susan wanted to buy the (new)^{T} [one]$^{T}_{2}$.*
(19) *John preferred (the)^{T} (old)^{C} [car]$^{T}_{1}$. Susan wanted to buy [it]$^{T}_{1}$.*

The removal of the modifiers “old” and “new” in (18) would make both referring expressions to refer to the same car (“John preferred the car. However, Susan wanted to buy the car”). Therefore the modifiers are required for determining the referent and so are distinctive. Removing the modifier “old” in (19) (“John preferred the car. Susan wanted to buy it”) would make the text no less understandable, and would not change what either referring expression “points to.” Therefore “old” is descriptive in (19). Thus rule #7: **If modifiers can be removed without changing the assumed referent of a referring expression, the modifier is descriptive.**

Sometimes a constituent is needed to identify the referent of the referring expression, but is said with an unusual or flowery choice of words that itself is descriptive. Consider:

(20) *John liked (the)^{T} (white)^{C} [fox terrier]^{C}. Mary liked it too.*

In (20), the modifier is again giving us additional information that we don’t need for distinction, and so it is again descriptive. The nucleus, however, describes the dog as a “fox terrier.” This is more specific than needed, and more specific than people normally speak, so the nucleus is descriptive. This is one of the key differences between a descriptive and distinctive referring expression. We have normal ways of talking about things: we usually refer to dogs as dogs, not as animals, mammals, or half-Doberman-half-Maltese, *unless* we have a specific piece of information we are trying to communicate (or purpose we are trying to achieve) above and beyond identifying the referent of the expression. To do this, you have to try and “get into the head” of the author, and imagine what are the other ways the referring expression could be expressed. We try to capture this in rule #8: **If the choice of words for a modifier or nuclei goes above and beyond the normal phrases used in such a context, it is probably descriptive**.

As you can see, the difference between distinctive and descriptive is subtle and sometimes a bit subjective. It has to do with at least the following four items:

1. The other things being talked about in the immediate context, both before and after the referring expression in question
2. The other things in the broader world that might be talked about
3. The way the information is normally presented: modifier and nuclei word choice

4. The way the information has already been presented.

Assigning a function to nuclei and modifiers can be a tricky task. One rule that can be used for helping in the annotation is rule #9: **First determine if the referring expression as a whole is distinctive or descriptive. If the referring expression is descriptive, it must contain at least one descriptive nucleus or descriptive modifier.**

Annotating copular predicates

The final piece of information to include in an indication is whether the referent makes up the predicate part of the copular expression. This is the ‘Y’ in an expression of the form “X is Y.” Keep in mind that an expression is a copular expression when the verb between the X and the Y is a linking verb, and can be any tense. A linking verb is a verb that connects the subject X to more information about that subject (Y). A good rule of thumb is if you can substitute a form of the verb “be” in its place (is, are, am, was, were, has been, are being, might have been, etc.), and the sentence still makes sense, it is probably a linking verb.

Common Cases

This section covers common cases which have a standard interpretation.

Sub-Referring expressions

Many referring expressions contain other referring expressions (their sub-referring expressions). Make sure you don’t split a sub-referring expression between multiple constituents. For example, in (21), the referring expression “the owner of the orchard” does not split its sub-referring expression “an orchard” among multiple constituents – “an orchard” is fully contained in a single modifier.

Determiners

Determiners (articles and demonstratives) are almost always distinctive modifiers. They should be put in a modifier by themselves, when not more closely attached to another modifiers (or sub-referring expression). Consider the following example:

(21) *Antonius was <u>(the)^T [owner]^T (of an orchard)^T</u>.*

Here the first “the” is modifying “owner”, and is distinctive. The second determiner, “an”, attaches to “orchard”, and so is part of the modifier “of an orchard.” “An orchard” is itself a referring expression, and it would be broken up like so:

(22) *Antonius was the owner of <u>(an)^T [orchard]^T</u>.*

Pronouns

Pronouns, when found alone, are usually distinctive, especially when referring to people. However, there are cases when they can be descriptive. The most common case is when we do not know the gender of the referent, and it is introduced by the use of male or female pronoun:

(23) John had a beautiful dog. <u>[She]</u>^{C} was the most beloved dog of the neighborhood.

In (23) the pronoun “she” is providing the extra information of the sex of the dog, so it is also descriptive. In this case “it” would have been enough.

Personal Titles

Personal titles should almost always be distinctive, and are usually their own modifiers:

(24) *(Mr.)^T [Smith]^T goes to Hollywood.*

Quantities & Numbers

Words and phrases indicating quantities like “most of,” “some of,” and “a few” should be kept together and should usually be marked as distinctive modifiers. Similarly, numbers are almost always their own distinctive modifier:

(25) *(many of)^T (the)^T (25)^T [countries]^T*

Symbols

Symbols such as percentages and currency markers should be included with the constituent they are most closely associated with, and **not** made into a separate constituent.

(26) *It will cost [7.5 %] (more per subscriber).*
(27) *It will cost [$100,980].*

Dates

When annotating the dates, you must keep in mind indication rule #3, that the basic type of the nucleus should be the same as the referring expression as a whole. If the intention of the author is to refer to a day, then the numeral portion of a date such as “the 30th of April” should be marked as the nucleus, and the month as a modifier, but only if indication rule #7 can be followed (28). If the referring expression is “April 30,” as in (29), it must be marked as a whole nucleus, since you cannot drop out the token “April” without changing the class of the referent. If the intention is a month, for example, “April 2010”, the month is the nucleus and the year the modifier (30). However if the date is being used itself as a modifier, the date should be kept together in accordance with our injunction against splitting sub-referring expressions across constituents (31).

(28) *He will measure the performance of these countries by (the) [30th] (of April).*
(29) *He will measure the performance of these countries by [April 30].*
(30) *He will measure the performance of these countries by [April] (2010).*
(31) *In your (Oct. 6) [article], you mention…*

Plural Referring Expressions

While you are allowed to mark multiple nuclei for a referring expression (and the nuclei are allowed to be discontinuous), a referring expression will almost always have a single nucleus. This is especially important to keep in mind when dealing with plural referring expressions:

(32) *[[Hansel] and [Gretel]] slept.*

Here there are three referring expressions: “Hansel”, “Gretel”, and “Hansel and Gretel.” The first two have only a single token, and so their nuclei are the whole referring expression. The last is a plural referring expression, the conjunction of two other referring expressions. Since the basic type of the nucleus must be the same as the referring expression as a whole, the plural referring expression’s nucleus must contain all the tokens.

Difficult Cases

“X of Y” expressions

In certain cases it is not clear which words are the nuclei and which the modifiers. Consider the following underlined referring expression:

(33) *The city of Boston has elections tomorrow.*

Is “city” the nucleus and “of Boston” the modifier? Or is “Boston” the nucleus and “city of” the modifier? In these cases we must remember that the nucleus is the part of the referring expression that performs the core referential function. We can often (but not always) reveal this by varying the referring expression and seeing what happens. For example, the substitution of the referring expression in (33) by “the city” could be ambiguous (or strange, or ill-formed) in some contexts, but if it is substituted by “Boston” would have the same meaning and referent. In those cases “Boston” would be the nucleus of the indication and “the city of” is the modifier.

In some cases, either substitution will be acceptable. In those cases, choose the most specific noun as the nucleus if it is otherwise undecidable**.** If the case above were still unclear after varying the referring expression, we would choose “Boston” as the nucleus because it is more specific than “city.”

Summary of Rules

Referring Expressions & Coreference

#	Rule
1	Mark noun phrases as referring expressions
2	Mark referential expressions that corefer with marked referential expressions, even if they would not normally be marked
3	Mark generics as referring expressions only when they are referred to directly
4	Numerals and pronouns should usually be marked as referring expressions
5	Mark quantifiers, determiners, adjectives, adjectival phrases, prepositional phrases, and other modifiers as part of the referring expression
6	With quantified referring expressions, use variation of quantifiers to test coreference
7	Use the substitution test to determine appropriate coreference relations.

Indications

#	Rule
1	Mark the core referential words of each referring expression as the nucleus (usually nouns, and usually the heads of the phrase)
2	Most references have one and only one nucleus
3	The nucleus should have the same basic type as the referring expression itself
4	Exclude as many words as possible from the nucleus
5	Most non-nucleus words, no matter their parts of speech, will be part of some modifier
6	Mark a nucleus or modifier as descriptive if it contains more information than needed to unambiguously identify the referent of the referring expression
7	If modifiers can be removed without changing the assumed referent of a referring expression, the modifier is descriptive
8	If the choice of words for a modifier or nuclei goes above and beyond the normal phrases used in such a context, it is probably descriptive
9	First determine if the referring expression as a whole is distinctive or descriptive. If the reference is descriptive, it must contain at least one descriptive nucleus or descriptive modifier

Glossary

appositive a noun or noun phrase that describes another noun or noun phrase directly adjacent.

copular expression A sentence (or phrase) of the form "X v Y," where X is the subject, Y the object, and v is a linking verb.

copular predicate The object in a copular expression.

coreferential The relationship that holds between two references when they refer to the same referent.

linking verb Connects the subject being described with a description.

metonymy see *synecdoche.*

referring expression A set of words that indicates a referent. For every referent mentioned in a text there may be multiple references.

referent Something that we talk about. Referents may be concrete or abstract, real or imagined; they may be objects, times, quantities, events, or any number of other things.

synecdoche (a.k.a. metonymy) a figure of speech in which a part of an object, or a closely related object, is used to refer to the whole object.