Coreference

In linguistics, **coreference**, sometimes written **co-reference**, occurs when two or more expressions in a text refer to the same person or thing; they have the same referent, e.g. *Bill said he would come*; the proper noun *Bill* and the pronoun *he* refer to the same person, namely to Bill.[1] Coreference is the main concept underlying binding phenomena in the field of syntax. The theory of binding explores the syntactic relationship that exists between coreferential expressions in sentences and texts. When two expressions are coreferential, the one is usually a full form (the antecedent) and the other is an abbreviated form (a proform or anaphor). Linguists use indices to show coreference, as with the i index in the example $Bill_i$ *said* he_i *would come*. The two expressions with the same reference are *coindexed*, hence in this example *Bill* and *he* are coindexed, indicating that they should be interpreted as coreferential.

Contents

Types

When exploring coreference, there are numerous distinctions that can be made, e.g. anaphora, cataphora, split antecedents, coreferring noun phrases, etc.[2] When dealing with proforms (pronouns, pro-verbs, pro-adjectives, etc.), one distinguishes between anaphora and cataphora. When the proform follows the expression to which it refers, anaphora is present (the proform is an anaphor), and when it precedes the expression to which it refers, cataphora is present (the proform is a cataphor). These notions are illustrated as follows:

Anaphora

a. **The music$_i$** was so loud that **it$_i$** couldn't be enjoyed. –The anaphor *it* follows the expression to which it refers (its antecedent).

b. **Our neighbors$_i$** dislike the music. If **they$_i$** are angry, the cops will show up soon. – The anaphor *they* follows the expression to which it refers (its antecedent).

Cataphora

a. If **they$_i$** are angry about the music, **the neighbors$_i$** will call the cops. – The cataphor *they* precedes the expression to which it refers (its postcedent).

b. Despite **her$_i$** difficulty, **Wilma$_i$** came to understand the point. – The cataphor *her* precedes the expression to which it refers (its postcedent)

Split antecedents

a. **Carol$_i$** told **Bob$_j$** to attend the party. **They$_{i+j}$** arrived together. – The anaphor *they* has a split antecedent, referring to both *Carol* and *Bob*.

b. When **Carol$_i$** helps **Bob$_j$** and **Bob$_j$** helps **Carol$_i$**, **they$_{i+j}$** can accomplish any task. – The anaphor *they* has a split antecedent, referring to both *Carol* and *Bob*.

Coreferring noun phrases

a. **The project leader$_i$** is refusing to help. **The jerk$_i$** thinks only of himself. – Coreferring noun phrases, whereby the second noun phrase is a predication over the first.

b. **Some of our colleagues$_1$** are going to be supportive. **These kinds of people$_1$** will earn our gratitude. – Coreferring noun phrases, whereby the second noun phrase is a predication over the first.

Versus bound variables

Semanticists and logicians sometimes draw a distinction between coreference and what is known as a bound variable.[3] An instance of a bound variable can look like coreference, but from a technical standpoint, one can argue that it actually is not. Bound variables occur when the antecedent to the proform is an indefinite quantified expression, e.g.[4]

a. **Every student$_i$** has received **his$_i$** grade. – The pronoun *his* is an example of a bound variable

b. **No student$_i$** was upset with **his$_i$** grade. – The pronoun *his* is an example of a bound variable

Quantified expressions such as *every student* and *no student* are, from a technical standpoint, not referential. The subjects *every student* and *no student* are grammatically singular, but they do not pick out single referents in the discourse world. Thus since the antecedents to the possessive adjective *his* is not referential, one also cannot say that *his* is referential. Instead, one says it is a *variable* that is *bound* by its antecedent. Its reference varies based upon which of the students in the discourse world is thought of. If Jack, John, and Jerry are the three students in the discourse world, then the meaning of *his* varies based upon whether Jack, John, or Jerry is the focus of the minds eye. The existence of bound variables is perhaps more apparent with the following example:

c. **Only Jack$_i$** likes **his$_i$** grade. – The pronoun *his* can be a bound variable.

This sentence is ambiguous. It can mean that Jack likes his grade, but everyone else dislikes Jack's grade, or more likely, it means that Jack likes his grade, but John dislikes his (John's) grade, and Jerry dislikes his (Jerry's) grade. The second, more natural reading is the bound variable reading. While the distinction between coreference and bound variables may be real, coindexation can be construed as accommodating both. That is, when two or more expressions are coindexed, it indicates that one is dealing with coreference or a bound variable.

Coreference resolution

In computational linguistics, coreference resolution is a well-studied problem in discourse. To derive the correct interpretation of a text, or even to estimate the relative importance of various mentioned subjects, pronouns and other referring expressions must be connected to the right individuals. Algorithms intended to resolve coreferences commonly look first for the nearest preceding individual that is compatible with the referring expression. For example, *she* might attach to a preceding expression such as *the woman* or *Anne*, but not to *Bill*. Pronouns such as *himself* have much stricter constraints. Algorithms for resolving coreference tend to have accuracy in the 75% range. As with many linguistic tasks, there is a tradeoff between precision and recall.

A classic problem for coreference resolution in English is the pronoun *it*, which has many uses. *It* can refer much like *he* and *she*, except that it generally refers to inanimate objects (the rules are actually more complex: animals may be any of *it*, *he*, or *she*; ships are traditionally *she*; hurricanes are usually *it* despite having gendered names). *It* can also refer to abstractions rather than beings: "He was paid minimum wage, but didn't seem to mind it." Finally, *it* also has pleonastic uses, which do not refer to anything specific:

a. **It**'s raining.

b. **It**'s really a shame.

c. **It** takes a lot of work to succeed.

d. Sometimes **it**'s the loudest who have the most influence.

Pleonastic uses are not considered referential, and so are not part of coreference.[5]

See also

- Anaphora (linguistics)
- Antecedent
- Binding
- Cataphora
- Nearest referent
- Switch reference
- Word-sense disambiguation

Notes

1. For definitions of coreference, see for instance Crystal (1997:94) and Radford (2004:332).
2. These distinctions (anaphora, cataphora, split antecedents, coreferring noun phrases, etc.) are discussed in Jurafsky and Martin (2000:669ff.).
3. For discussions of bound variables, see for instance Portner (2005:102ff.).
4. See Jurafsky and Martin (2000:701) for an example of a bound variable like the ones given here.
5. Li et al. (2009) have demonstrated high accuracy in sorting out pleonastic *it*, and this success promises to improve the accuracy of coreference resolution overall.

References

- Crystal, D. 1997. A dictionary of linguistics and phonetics. 4th edition. Cambridge, MA: Blackwell Publishing.
- Jurafsky, D. and H. Martin 2000. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. New Delhi, India: Pearson Education.
- Portner, P. 2005. What is semantics?: Fundamentals of formal semantics. Malden, MA: Blackwell Publishing.
- Radford, A. 2004. English syntax: An introduction (https://books.google.com/books?id=LdAi292Q4-0C&printsec=frontcover#v=onepage&q=coreference&f=false). Cambridge, UK: Cambridge University Press.
- Yifan, L., P. Musilek, M. Reformat, and L. Wyard-Scott 2009. Identification of pleonastic *it* using the web (http://www.jair.org/media/2622/live-2622-4361-jair.pdf). *Journal of Artificial Intelligence Research* 34, 339–389.