1

What Is a Semantic Annotation?

1.1 Annotation: Past and Present

1.1.1 Traditional Scholarship

Annotation literally means *adding notes* to text or images. Like commentary work, it is scholarly work with a long historical tradition. It has specific methodological merits for describing or explaining what has been given to scholars or teachers of classical Greek or Latin literature, biblical exegetists of the Hebrew Bible, philosophers of Chinese writings or monks of Buddhist sutras. They have thus produced scholarly books such as *The Aeneid Annotated Virgil*,[1] *Cambridge Annotated Study Bible*,[2] as shown in Figure 1.1, *The New Oxford Annotated Bible*,[3] *A New Translation of Lunyu with Annotations*,[4] or *The Diamond Prajna-Paramita Sutra (The Diamond Sutra): An Annotated Edition with Chinese Text*.[5]

Some people think of annotation as an outdated business or archaic scholarly methodology. You pick up a short list of terms and sometimes make nothing but a lengthy unconnected series of commentaries on those terms, as is sometimes complained. Just as linguists are often understood as polyglots, those who work on annotation would be considered as treating ancient texts or things of antiquities only. Adding notes has, however, been taken as a serious scholarly work through the ages. Figure 1.2 shows that a grammar book was written with *critical notes*.

[1] By Virgil. Translated by John Dryden, Kindle Edition.
[2] Edited by Howard Kee, Cambridge University Press, 1993.
[3] Edited by Bruce M. Metzer and Roland E. Murphy, New York: Oxford University Press, 1991,1994.
[4] This is a subtitle for the book *Understanding the Analects of Confucius* by Peimin Ni, Albany, NY: State University of New York Press, March 2017.
[5] Translated and annotated by Ven. Cheng Kuan, 2nd ed., 2017, American Buddhist Temple, USA.

https://doi.org/10.1017/9781108884532.004 Published online by Cambridge University Press

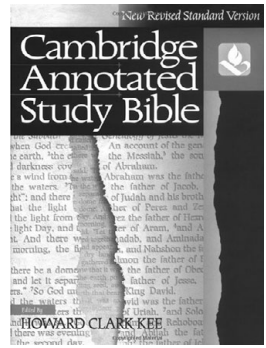

Figure 1.1 Annotated Bible
Reprinted by permission from Cambridge University Press.
Kee, Howard C. (1993) *Cambridge Annotated Study Bible*.

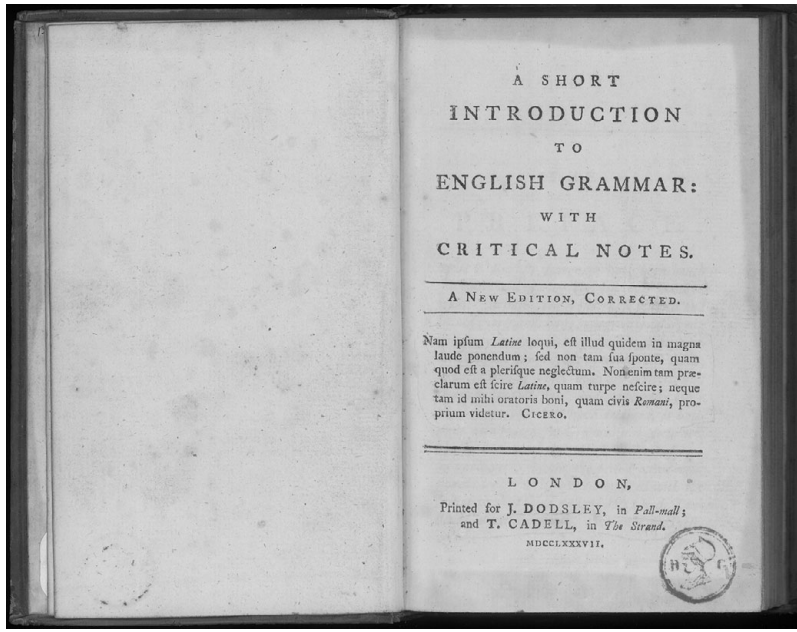

Figure 1.2 Grammar with critical notes
Ghent University Library, BIB.BL.000976.

Annotation is an activity with products that are also called *annotations*. It enriches the main content of a text. It resolves lexical or sentential ambiguities, provides underspecified textual meanings with contextual or background information, and updates described situations that are either diachronically outdated

https://doi.org/10.1017/9781108884532.004 Published online by Cambridge University Press

or synchronically outplaced by introducing relevant explanatory information. Formats have thus been developed to represent a variety of information added to the main text.

1.1.2 Formats for Annotation

There are at least four commonly accepted ways of adding notes to the main text: *innotes, footnotes, sidenotes*, and *endnotes*. Innotes are inserted into the main content part of a text with parentheses, especially when notes are brief. Innotes can take up a good portion of the main part of a page, for instance, either by alternating a paragraph with the main content and the following paragraph with commenting notes or by occupying a column within or next to the main part.

Cambridge's annotated Bible contains footnotes at the bottom of a page and two columns of sidenotes on the left and right sides of the page. These notes have different uses, as shown in Figure 1.3.

There are two footnotes at the bottom of the main text in Figure 1.3. They are each linked by an alphabet letter *a* and *b* to the term which is being annotated, as shown by the two arrows. The sidenotes on the left side column are references to citations in the Bible that are related to the verse under discussion, whereas the sidenotes on the right side column contain comments on the verse, or the sequence of verses of the chapter referred to.

Endnotes are listed at the end of a chapter or a book, again being referred to by a number to the annotated term. Whatever format for notes there might have been, all these notes were included by chapter in a volume that carries its main content as a book.

In modern times, the way of providing additional information has become more sophisticated as the technology of printing and photography has developed. The task of adding extra information is carried out by relevant illustrations or photos of varying data to the degree that these visualizations are considered part of the main textual content. *The Cambridge Encyclopedia of the English Language* is a good example (Figure 1.4).The page contains three notes: two notes on the right column of the page and a third one from the previous page linking to a map with several arrows showing the origins of English. The map is a part of the third note.

How to lay out additional information and what to introduce as additional information are issues that are constantly asked. Such questions are seriously taken up when the text turns into electronically manageable files or datasets for the merging, interchange, and evaluation of information in them. A variety

https://doi.org/10.1017/9781108884532.004 Published online by Cambridge University Press

GENESIS

Six Days of Creation and the Sabbath

1 In the beginning when God created[a] the heavens and the earth, [2]the earth was a formless void and darkness covered the face of the deep, while a wind from God[b] swept over the face of the waters. [3]Then God said, "Let there be light"; and there was light. [4]And God saw that the light was good; and God separated the light from the darkness. [5]God called the light Day, and the darkness he called Night. And there was evening and there was morning, the first day.

6 And God said, "Let there be a dome in the midst of the waters, and let it separate the waters from the waters." [7]So God made the dome and separated the waters that were under the dome from the waters that were above the dome. And it was so. [8]God called the dome Sky. And there was evening and there was morning, the second day.

9 And God said, "Let the waters under the sky be gathered together into one place, and let the dry land appear." And it was so. [10]God called the dry land Earth, and the waters that were gathered together he called Seas. And God saw that it was good. [11]Then God said, "Let the earth put forth vegetation: plants yielding seed, and fruit trees of every kind on earth that bear fruit with the seed in it." And it was so. [12]The earth brought forth vegetation: plants yielding seed of every kind, and trees of every kind bearing fruit with the seed in it. And God saw that it was good. [13]And there was evening and there was morning, the third day.

14 And God said, "Let there be lights in the dome of the sky to separate the day from the night; and let them be for signs and for seasons and for days and years, [15]and let them be lights in the dome of the sky to give light upon the earth." And it was so. [16]God made the two great lights—the greater light to rule the day and the lesser light to rule the night—and the stars. [17]God set them in the dome of the sky to give light upon the earth, [18]to rule over the day and over the night, and to separate the light from the darkness. And God saw that it was good. [19]And there was evening and there was morning, the fourth day.

20 And God said, "Let the waters bring forth swarms of living creatures, and let birds fly above the earth across the dome of the sky." [21]So God created the great sea monsters and every living creature that moves, of every kind, with which the waters swarm, and every winged bird of every kind. And God saw that it was good. [22]God blessed them, saying, "Be fruitful and multiply and fill the waters in the seas, and let birds multiply on the earth." [23]And there was evening and there was morning, the fifth day.

24 And God said, "Let the earth bring forth living creatures of every kind: cattle and creeping things and wild animals of the earth of every kind." And it was so. [25]God made the wild animals of the earth of every kind, and the cattle of every kind, and everything that creeps upon the ground of every kind. And God saw that it was good.

26 Then God said, "Let us make humankind[c] in our image, according to our likeness; and let them have dominion over the fish of the sea, and over the birds of the air, and over the cattle, and over all the wild animals of the earth,[c] and over every creeping thing that creeps upon the earth."

27 So God created humankind[d] in his image,
in the image of God he created them;[e]
male and female he created them.

[28]God blessed them, and God said to them, "Be fruitful and multiply, and fill the earth and subdue it; and have dominion over the fish of the sea and over the birds of the air and over every living thing that moves upon the earth." [29]God said, "See, I have given you every plant yielding seed that is upon the face of all the earth, and every tree with seed in its fruit; you shall have them for food. [30]And to every beast of the earth, and to every bird of the air, and to everything that creeps on the earth, everything that has the breath of life, I have given every green plant for food." And it was so. [31]God saw everything that he had made, and indeed, it was very good. And there was evening and there was morning, the sixth day.

[a] Or *when God began to create* or *In the beginning God created* [b] Or *while the spirit of God* or *while a mighty wind* [c] Heb *adam* [d] Syr Heb *and over all the earth* [e] Heb *him*

1

See the Introductions, pp. 2, 30, and 32-33 above.

1.1-2.4a The Priestly Account of the Creation. The emphasis falls on the sovereignty of God and the orderliness of the process of creation. Throughout this section, God is given the name *elohim* in the Hebrew original.

1.2 *The earth was a formless void.* God forms and orders the world out of existing, chaotic matter. *The deep...the waters.* In the mythology of Canaan and Mesopotamia the waters were the symbols of chaos which the more powerful beneficent deities had to bring under control.

1.3 *God said.* The power of God to achieve his purpose is evident when he speaks his intention and it is accomplished.

1.4 *God separated the light from the darkness.* The ordering of light and darkness establishes the rhythm of time, with evening followed by morning, which is the principle of Israelite days beginning at sundown.

1.6-10 *God...separated the waters...the waters that were gathered together* God's ordering of the world results in the separation of sky and earth, of sea and dry land.

1.14-19 The ordering of day and night is accomplished by the positioning of the sun, moon and stars

1.26-28 *Let us make humankind in our image.* The Hebrew word for man is *adam*, which serves as the name of the first human being in these creation stories. Essential to the role of humans created in God's image is their exercise of authority over the earth and all living things upon it.

1.1 Jn 1.1,2; Ps 8.3; Isa 44 24; 42 5; 45.18

1.2 Jer 4.23, Ps 104.30

1.3 Ps 33.6,9, 2 Cor 4.6

1.4 Isa 45.7

1.5 Ps 74.16

1.6 Jer 10.12

1.7 Prov 8.28, Ps 148 4

1.9 Job 26.10, Prov 8.29; Jer 5.22, 2 Pet 3 5

1.10 Ps 33 7

1.11 Lk 6 44

1.14 Ps 74 16, 104 19

1.16 Ps 136.8,9; Job 38 7

1.18 Jer 31 35

1.21 Ps 104.25,26

1.22 Gen 8.17

1.25 Jer 27.5

1.26 Ps 100.3; Acts 17 26; Col 3.10

1.27 1 Cor 11.7; Gen 5.2; Mt 19.4

1.28 Gen 9.1,7; Lev 26.9

Figure 1.3 Genesis annotated
Reprinted by permission from Cambridge University Press.
Kee, Howard C. (1993) *Cambridge Annotated Study Bible*.

of formats representing annotation have been proposed in the area of computational work, including tabular formats with vertical columns and graphs.

1.1.3 Taking a New Turn

With the advance of the age of information and computation, the status of annotation has changed as it applies to the analysis of human natural language

https://doi.org/10.1017/9781108884532.004 Published online by Cambridge University Press

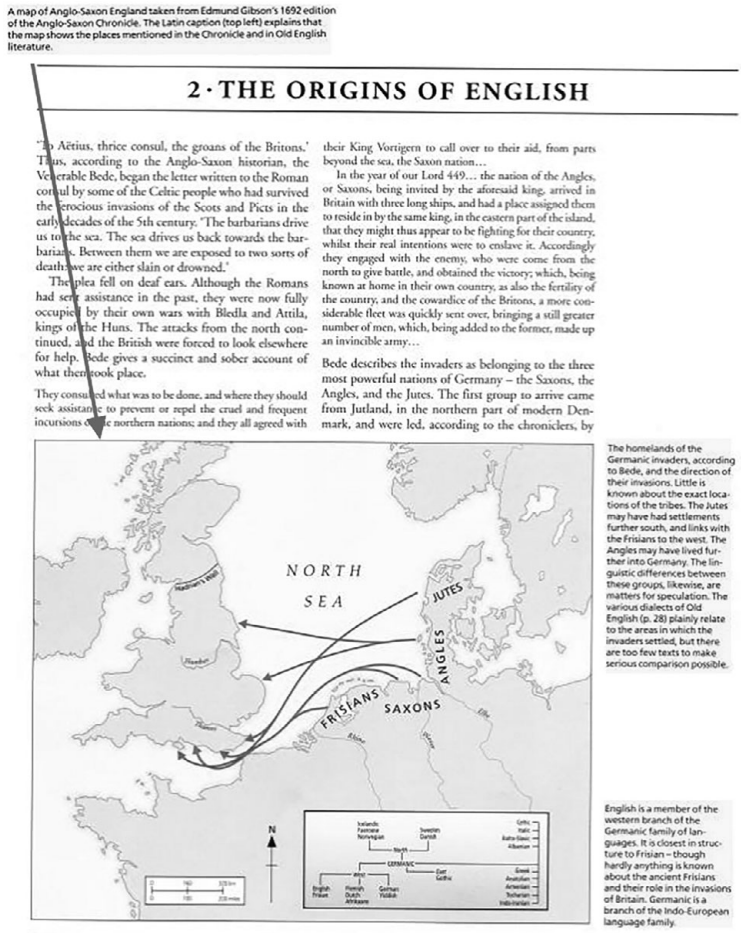

Figure 1.4 Visual illustration for additional information
Reprinted by permission from Cambridge University Press.
Crystal, D. (2003) *The Cambridge Encyclopedia of the English Language*.

rendered in various forms, whether written, spoken, or visualized as static or dynamic images (pictures, photos, or videos). Being subject to computational processing, *text* no longer refers to a simple collection of fragments of written material or printed matter, but a computationally readable file that carries information or messages to convey. Likewise, text messaging or texting refers to the activity of composing and sending electronic messages. The annotation of such text is now an essential part of the field of natural language processing

https://doi.org/10.1017/9781108884532.004 Published online by Cambridge University Press

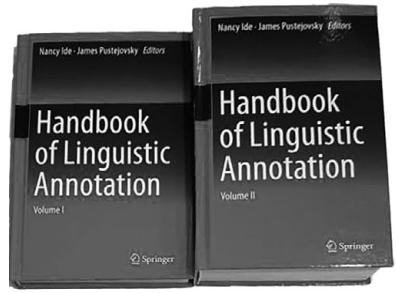

Figure 1.5 *Handbook of Linguistic Annotation*
Reprinted by permission from Springer Nature, Ide, N., and Pustejovsky, J. (eds.) *Handbook of Linguistic Annotation*, Volumes 1 and 2, Springer, Berlin, Heidelberg © 2017.

(NLP) with its scientific technology, now called *linguistic annotation*, as witnessed by the appearance of the two-volume *Handbook of Linguistic Annotation* (see Figure 1.5). Linguistic annotation is the basis of NLP.

1.2 Linguistic Annotation

1.2.1 Overview

In the 1960s, linguistic annotation started with the building of large amounts of language data, called *corpus linguistics*. The time of its beginning was not very favorable. First, the research or academic environment for any statistical work was poorly developed. This was especially so because most of the linguists, especially in America, were fascinated with Chomsky's theory of generative grammar that focused on the so-called ideal speaker's intuitive judgments on language facts. This theory may have succeeded in deepening the psychological understanding of how the human faculty works in the use of language, while ignoring the practical limitations of human cognition and linguistic performance. It also underestimated the statistical power of predicting human interactions in communication. Faced with complex issues or even a simple but deeply iterative structure, the performance of human capacity rapidly fails to function reliably. When even well-trained linguistics students are asked to evaluate the well-formedness of strings of words as grammatically correct sentences, they quickly become tired of making a valid and reliable judgment, especially if those strings are repeatedly read out to them or if they are coerced to make a decision.

Second, no materials or tools were easily available. Computer-readable material was almost nil at that time. Personal desktop computers came out around the 1980s. When portable laptops such as Apple or IBM XT were made available, ordinary office workers with no linguistics background were hired to type in text manually to convert it to electronic files. Books and newspapers had not been published electronically. Furthermore, there were no standardized coding systems like the American Standard Code for Information Interchange (ASCII)[6] or Unicode (the Unicode Standard for the Universal Character Set).[7]

Despite all these difficulties, corpus linguistics has now come into the mainstream of linguistics. It has been established not so much as an independent part of general linguistics, but more so as a fundamental methodology applicable to the whole range of linguistics from phonology to morpho-syntax to semantics, pragmatics, and discourse analysis, as well as to the new area of computational linguistics in particular. Here textual annotation forms a basic framework for applying such a methodology to the processing of datasets in language. Tagging, markup, and parsing are kinds of annotation in NLP, each of which applies to the building of so-called *annotated corpora* by providing extratextual information, called *metadata*, to a given dataset.

Large Data From the Internet, we can now quickly obtain a large amount of data from natural language. News items, research articles, maps and pictures, and all other sorts of information in various domains are easily accessible through Wikipedia, Google Maps, Google Search, Research Gate, or ChatGPT. Promotional emails also pour out a lot of information. All of them are now electronically manageable, providing possible data that can be built into a corpus only if some legal barriers such as copyright or privacy laws are resolved.

[6] ASCII defines all of the 26 alphabet letters, called Latin characters, in upper or lower case, assigning a unique code point to each of them in the set of 128 character code points represented in 7 bits from 0000000 to 1111111. The capital (upper-case) letter "A", for instance, is represented by 1000001 in binary. The first edition was published in 1963 and the latest edition in 1986, mainly under the American National Standards Institute (ANSI), an active member of the International Organization for Standardization (ISO).

[7] The first draft proposal, called *Unicode*, came out in August 1988 for an international or multilingual text character coding system. The first version of *the Unicode Standard* was published in 1991, and now version 12.0.0 is available by the Unicode Consortium. ASCII was incorporated into Unicode. Lacking a unified coding system, it was impossible to combine various electronic files to build a very large collection of data, which could be genuinely called a *corpus*, in a consistently efficient way. This had been the case with corpora, especially in languages that used non-Latin alphabet characters (Graham, 2000).

1.2.2 Kinds of Tasks

Given some language data, it is segmented into characters or strings of character segments called *tokens*. These tokens are then grouped to form larger strings of characters, called *words*, and each of these words is classified with a morpho-syntactic category such as a noun or a verb. They are also grouped into larger units, called *phrases* or *chunks*, again with appropriate category names. The addition of such category names to a given dataset provides extra information which we have been calling *metadata*. Such segmentation or grouping allows the identification of portions of text or images, called *markables* for annotation. Strictly speaking, such tasks are not part of annotation, but a necessary step of processing primary data before identifying markables for annotation. Annotation, applied to NLP, means not just adding plain notes, but very often adding lexical information with the names of syntactic categories to segmented data. Such work is the most typical sort of corpus annotation, called *part-of-speech (POS) tagging*, contributing to the resolution of lexical or structural ambiguities contained in input phrases or words. Here is a well-known ambiguous sentence, called a *garden path sentence*.

Example 1.1 POS-tagging a garden path sentence

a. The horse raced past the barn fell.
b. The horse raced$_{VVD}$ past the barn fell. (fails to be processed)
c. The horse raced$_{VVN}$ past the barn fell. (succeeds in being processed)

The tagging of a word *raced* as VVD (past-tense verb) fails to process Example 1.1a when the processing step reaches the verb *fell*. In contrast, with the tagging of the word *raced* as VVN (past participle), Example 1.1a is successfully processed, as annotated in 1.2.[8]

Annotation 1.2 Annotating the garden path sentence

The horse [that was raced$_{VVN:past\ participle}$ [past$_{PRP:preposition}$ the barn]] fell$_{VVD:past\ tense}$.

Such a task of tagging words with grammatical categories or class names is a proper part of the annotation. It is, however, treated as a preprocessing step for semantic annotation.

Named entity disambiguation (NED) is, in contrast, considered part of semantic annotation. For example, the string of three words *the White House* refers typically to the official residence and workplace of the US President,

[8] The grammatical tags VVD and VVN are taken from the British National Corpus (BNC) Basic (C5) tagset. They stand for the past tense form of lexical verbs (e.g., *forgot, sent, lived, returned*) and the past participle form of lexical verbs (e.g., *forgotten, sent, lived, returned*), respectively.

https://doi.org/10.1017/9781108884532.004 Published online by Cambridge University Press

but sometimes refers to its function as a metonymic expression. Here is a newspaper headline, which illustrates how the words *White House* are used.

Example 1.3 Newspaper headline
WHITE HOUSE ANNOUNCES TRUMP TO VISIT
SOUTHERN BORDER

The annotation of named entities such as one referred to by "WHITE HOUSE" provides different ways of annotating them; for example, as follows.

Annotation 1.4 Named entity disambiguation (NED)
`White House_<facility OR institution>`

The annotation of sentiments or metaphors may also be considered a proper part of annotation and also that of semantic annotation. Such an extension of annotation to language and its analysis requires highly developed technical training of humans and machines (computers) and also computer algorithms that require annotation structures as intermediate data structures for language processing.

1.2.3 Machine Learning

Machine learning theories are applied to natural language annotation to enhance its computational processing.[9] Base segmentation and subsequent tasks of tokenization and categorized chunking (see Chapter 2) as well as text mining for language resources are expected to be carried out by machines (see Figure 1.6).

Machine learning has become an essential topic in computational linguistics. The amount of data keeps increasing in various domains of interactive human languages through social networks or orally conveyed by dynamic human communications through television or communication applications like Skype or Zoom. Linguistic engineers thus find it necessary to be supported by machines or computers, which can run for 24 hours a day without complaining and breaking down, to process such data. Such data processing is ultimately required for the construction of practical systems for various NLP applications as well as various sorts of semantic annotation schemes for information encoding that supports such applications.

Humans train machines to annotate language. Humans form a group of annotation experts to prepare what and how to make machines learn by preparing a set of guidelines or norms, called *gold standards*. In preparing it,

[9] See two recent publications on annotation and machine learning: Pustejovsky and Stubbs (2012) and Meteer (2015).

https://doi.org/10.1017/9781108884532.004 Published online by Cambridge University Press

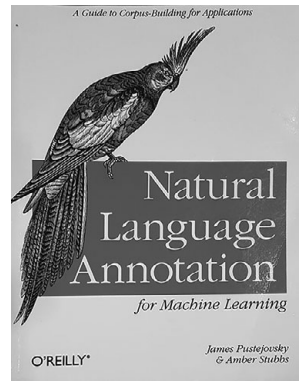

Figure 1.6 Machine learning to annotate language
Reprinted by permission from O'Reilly Media, Inc.
Pustejovsky, J., and Stubbs, A. (2012)
Natural Language Annotation for Machine Learning.

the annotation experts have to reach an agreement, called *interannotator agreement* (IAA), that guarantees the validity and reliability of human judgments on linguistic facts. The validity of IAA, very often measured statistically, supports the correctness of decisions, while the reliability retains the consistency of tasks on differing types of input data for annotation.

Making machines learn is not a simple one-step process. It requires a cycle of repeated but incremental steps of modeling (M) and annotating (A), possibly skipping the four additional steps: train (T), test (T), evaluation (E), and revision (R). The specification of annotation tasks itself needs to be revised continuously. Such a process is called MAMA by Pustejovsky and Stubbs (2012), which is depicted as a part of a longer process, called MATTER, in Figure 1.7.[10]

The process of MATTER consists of six steps in a cycle.

Specification 1.5 The development cycle of MATTER
(1) *Model* a given task to produce an annotation guidelines
(2) *Annotate* sample datasets
(3) *Train* human annotators and machine learners
(4) *Test* annotation results

[10] Refer to Pustejovsky and Stubbs (2012, Figure 1-10) for the basic concepts of the MAMA and MATTER cycles. The two inner cycles were added by the author (Kiyong Lee) of this book. MAMA refers to the inner-outer cycle (dotted line) of MATTER, but should also be referring to the innermost cycle, consisting of two steps, Model an algorithm (M) and Annotate (A), as has been pointed out by an anonymous reviewer.

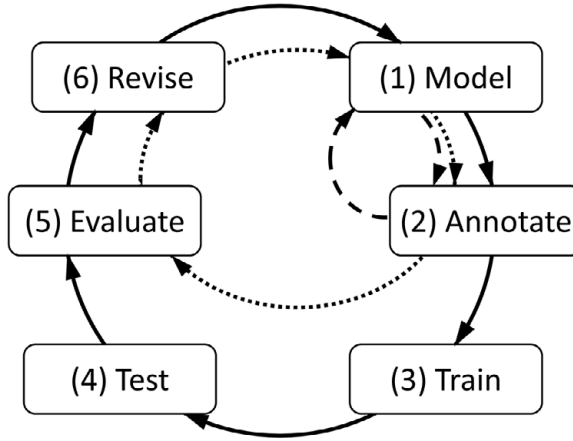

Figure 1.7 Process of training machine to learn
Reprinted by permission from O'Reilly Media, Inc.
Pustejovsky, J., and Stubbs, A. (2012)
Natural Language Annotation for Machine Learning.

(5) *Evaluate* interannotator agreement
(6) *Revise* annotation guidelines or algorithms

These six steps are connected by three cycles: the outermost solid line connecting all of the six steps (1) through (6), and the two inner cycles, one is a dashed line, and the other is a dotted line. The core portion of the figure is MAMA: the repeated cycle (innermost dashed line) of (1) *Model* (M) and (2) *Annotate* (A), supplemented by (5) *Evaluate* (E) and (6) *Revise* (R), as represented by the inner dotted line while bypassing the other two steps (4) *Train* (T) and (5) *Test* (T).[11] Model (M) an annotation task to specify an annotation guideline and then annotate (A). Repeat that process after the evaluation of annotation results and the revision of the annotation guideline. This process is repeated till satisfactory results are obtained.

Finlayson and Erjavec (2017) propose three additional stages, *Idea, Procure*, and *Distribute* for the MATTER and its inner cycle MAMA. The MATTER procedure starts with *Idea* that designs and formulates ideas and concepts for the project proposal, emphasizing the need for solid preparations. The step *Procure* means adopting a good tool for each of the well-defined subtasks of an annotation work with the concomitant belief that the selection of appropriate tools is as important as the designing of a good annotation scheme. The third

[11] For further details see Pustejovsky (2006) and Finlayson and Erjavec (2017).

https://doi.org/10.1017/9781108884532.004 Published online by Cambridge University Press

step *Distribute* is added to the end of the MATTER cycle, involving various techniques such as packaging, archiving, and exporting, to make the annotated resources available to the world at large.[12]

1.2.4 Levels of Linguistic Annotation

The traditional classification of linguistic levels was restricted to three areas.

Classification 1.6 Linguistic annotation levels
a. phonology
b. morphology
c. syntax

Phonology deals with patterns of sounds. It describes how sound segments, consonants, and vowels are classed into conceptually or perceptually identifiable *discrete units*, called *phonemes*, in a language, and also how strings of sound segments interact with each other in observationally systematic ways to be formulated as phonological rules of assimilation or ellipsis, etc. *Phonetics* that constructs the system of sounds, either human or physical, used to be treated as a preliminary part of phonology, just as logic was so for the philosophy that consists of metaphysics and epistemology, and so on.

Morphology defines words and their minimal units, called *morphemes*, and classifies words into grammatical categories, often called *parts of speech*, while formulating derivational rules for word formation. Finally, *syntax* formulates the rules of generation to form sentences out of words or sequences of words, called *phrases*, and to define (syntactically) *well-formed* sentences. If a sentence is well-formed with respect to a given set of rules (grammar), then it is said to be grammatical.

At the earlier stage of the history of linguistics, the separation of all these levels was strictly required. For instance, the classification of words in morphology should not depend on syntactic or semantic concepts. During the period of strict Structuralism that was prevalent in the 1930s and 1940s, Nouns and Verbs were thus named *Class 1* and *Class 2*, respectively, so that these class names were introduced as being independent of the semantic types of references expressed by the words that are being classified. Nouns and Verbs are, in contrast, derived from the meaning-bearing Latin words, *nomen* for nominal expressions (Noun) that refer to the names of entities and *verbum* for verbal expressions (Verb) that were the words referring to actions or states. The analysis of sentences should be described only in terms of structural

[12] See Figure 1 in Finlayson and Erjavec (2017).

https://doi.org/10.1017/9781108884532.004 Published online by Cambridge University Press

relations without making any reference to the propositional content carried by the sentences being analyzed. The grammatical function of *SubjectOf* is, for instance, defined *structurally* (in terms of structural relations) as a relation between the root of a phrase-structure tree (S) and its daughter node NP or any other category XP that precedes a node called VP (Verb Phrase) or *Predicate*.[13]

Syntax without semantics has now come to be considered useless or even unheard of. Since the 1970s, especially after the introduction of formal semantics like Montague Semantics, as introduced by Dowty et al. (1981), semantics has become an essential part of linguistics and such areas, called *pragmatics* and *analysis of dialogues and discourses*, have also been incorporated into the field of linguistics proper.

In linguistic annotation, the three levels of semantics, pragmatics, and analysis of dialogues and discourses have merged into one area, called *semantic annotation*. The part of morphology that treats word formation and lexical meanings is also incorporated into the semantic annotation. Phonetics is partially related to semantic annotation because some suprasegmental features of sounds such as stress, pitch, or loudness, and intonation patterns affect meaning in general and sentiments and moods in particular. Hence, all levels of linguistics, including multimodal aspects such as *gestures* or *facial expressions* involved in human communication, may be considered as contributing to semantic annotation.

Language is primarily spoken and contextually situated. Before videos were widely available, spoken data was transcribed, and transcribed data was then stored in a corpus. Phonetics and phonology provide scientific means for such transcription. Capturing visual information associated with linguistic data, especially related to human actions and motions of physical objects of intelligent agents, both humans and artificial robots, in time and space, has also become an essential part of linguistic annotation, especially for the contextually situated understanding of interactive human communication or the successful engineering of robotics. The semantic annotation should be contributing to such tasks beyond the treatment of ordinary text and moving towards treating all kinds of data involving multimodal communications that include gestures and facial expressions or the development of human machine interactions.

[13] This assumes that there is a phrase structure rule S → XP VP and that the *SubjectOf* relation is a relation [XP, S] according to Chomsky's theory of Generative Syntax. Note that XP is a generalization of NP to accommodate non-NP categories such as *From Seoul to Busan* as subjects in sentences like *From Seoul to Busan is approximately 500 km.*.

1.3 Semantic Annotation

1.3.1 Partial and Situated Information

Semantic annotation is characterized by the partiality and situatedness of information. These two characteristics form a theoretical basis for the modeling of semantic annotation schemes.

Partial Information

The range of markables for semantic annotation is restricted and specialized. It marks up relevant information from text or other media types of data in language that affects human actions only, focusing on some particular aspects of information in a restricted domain. The type and amount of information relevant for semantic annotation are thus very restrictive and partial, for ordinary human actions do not require so much information. Too much or overloaded information rather hinders the proper understanding of a given task, thereby deterring the proper performance of required appropriate actions. In general, semantic annotation works with a small world or a very tiny part of the spatio-temporally constrained world, but very seldom with the limitless universe of all possible worlds. It does not fit into *possible worlds semantics* that talks about the truth-condition or validity of propositions expressed by sentences uttered. Semantic annotation is thus focused on some particular aspects of a situation, viewed from some particular perspectives.

Semantic annotation does not mark up everything in a dataset, but selects a specific list of expressions, called *markables*, from the dataset which refers to certain types of entities. Event-oriented temporal annotation such as TimeML or ISO-TimeML selects those expressions as markables that refer to events and times as well as some time-related expressions such as temporal prepositions or conjunctions. Consider a short passage about the Appalachian Trail that runs from Georgia to Maine along the east coast of the United States. The expressions that are relevant for the question *When to start* are marked in *italics*.

Example 1.7 *When* should you *start* the Appalachian Trail?
The majority of thru-hikers *hikes northbound*, *beginning* in Georgia *anytime from late March to mid-April*. *Southbound hikers* generally *begin late May to mid-June*. Some hikers start heading north, then realize that they will not make it to Katahdin before Baxter State Park *closes on Oct. 15*.[14]

[14] Information from https://appalachiantrail.org/explore/hike-the-a-t/thru-hiking/northbound/, dated 2022-12-12.

https://doi.org/10.1017/9781108884532.004 Published online by Cambridge University Press

This passage provides an answer to the question of when to start hiking the Appalachian Trail. There are two possible directions for hiking: one is northbound and the other, southbound. A temporal annotation will focus on its markables, those expressions referring to events and times only, as listed in Annotation 1.8.

Annotation 1.8 Markables
a. hike northbound … begin … anytime from late March to mid-April
b. Southbound hikers … begin late May to mid-June
c. closes on Oct. 15

The last item which contains information about the closing time of Baxter State Park may be left out, for it simply provides background information about the reason why the northbound hike should start sometime in late spring.

Annotation thus focuses only on some relevant parts of the information that is provided by an input dataset without trying to capture all of the available pieces of information. Temporal annotation marks up only those expressions that refer to events and times and those signals that trigger relations over events and times. The annotation scheme will contain two types of markables, *event* and *time*, possibly with an extra type *signal*.[15]

There are, however, several or many different semantic annotation schemes with different foci, perspectives, and points of view, for many different types of information that are needed. The annotation scheme called TimeML, for instance, focuses on time and events, ISO-Space on locations, paths, and motions, or the annotation of semantic roles on participants in events. The integration or merging of all these different sorts of information calls for another task. If all these sorts of information were annotated simultaneously even for a short piece of text, it would take too much time to go through the whole annotation with the resulting annotation being too complicated to process and comprehend. However, if all these annotation schemes are built separately but desigend to be interoperable with each other, then there is no difficulty in merging them as the need arises.

Situated Information

The primary task of semantic annotation is to situate or put into context what has been described or uttered. This context can be a discourse situation in which something has been described or uttered, background information or belief that needs to be shared for successful dialogues, or any other type of situation that puts what needs to be interpreted into the right perspective.

[15] Signals have no referential status. They trigger some relations over entities and events.

https://doi.org/10.1017/9781108884532.004 Published online by Cambridge University Press

Suppose a traveler in the Berlin Hauptbahnhof is heading for Frankfurt and looking for a platform where she could hop on her train. She needs help, for the new Berlin Central Train Station is a huge place with 7 platforms and 16 tracks spread out to different destinations. So to be able to help her, one has to know a lot about the station but also where that particular traveler was standing when she asked for directions. The situation becomes more complex if the traveler is calling for someone through a mobile phone. The information provider may be a robot just standing where the traveler was standing or an intelligent phone system for travelers. These artificial agents are then helped by an intelligent interpreter based on some semantic annotation. All these agents need contextually situated information to act appropriately, as framed by Fillmore (1976).

Consider a short dialogue that involves another situation.

Example 1.9 Dialogue between speakers A and B
Speaker A: When did Mia leave for Boston?
Speaker B: At seven o'clock yesterday evening by Korean Airlines.

Speaker B gave the correct answer to A's question, but B's answer needs to be interpreted appropriately.

Ordinary semantics first reconstructs B's answer as a well-formed complete sentence like the following.

Example 1.10 B's answer reconstructed
Mia left for Boston at seven o'clock yesterday evening on Korean Airlines.

Only after some syntactic analysis, for instance, with Categorial Grammar, semantics starts interpreting each of the component phrases in the sentence by providing their meanings or intensions. The temporal expression *yesterday* is, for instance, interpreted as the 24 hours preceding the time of utterance. It is a the lexical meaning of *yesterday* that can be obtained from a lexicon.

Such an interpretation is not adequate for one who is going to wait for Mia's arrival in Boston. For her, the adequate temporal annotation will provide or compute the specific date and hour of Mia's departure by taking in various pieces of information relevant to the situation such as the time of utterance and the time zone difference between the place of Mia's departure and Boston. Annotation thus deals with *situated information* in such a specific way.

1.3.2 Tasks and Applications of Semantic Annotation

Semantic annotation marks up text or other forms of language data with various sorts of information that are necessary or relevant for performing

https://doi.org/10.1017/9781108884532.004 Published online by Cambridge University Press

communicative actions with the computer. Given computationally tractable datasets such as base-segmented or, more preferably, morpho-syntactically annotated data, semantic annotation enriches such data with information for high-level NLP applications that include information retrieval (IR), question-answering systems (QAS), machine translation (MT), text summarization, and spoken language understanding.

There are many different types of semantic annotation such as the annotation of word senses (e.g., various parallel corpora with the use of wordNet),[16] semantic roles (e.g., Frame Net, Propbank), time and events (e.g., TimeML, ISO-TimeML), locations and their qualitative spatial or directional relations (e.g., SpatialML), dialogue acts (e.g., DAMSL, DiAML) and discourse relation (e.g., Penn Discourse Treebank), and dynamic motions and transitions (e.g., ISO-Space). Each type of semantic annotation is characterized by its annotation scheme that defines a set of base categories and a set of links over base structures each based on a specific base category.

Illustrations A semantic annotation scheme for semantic role annotation specifies a set of two basic types, for instance, `<event>` and `<participant>`, and a link that relates an event to a participant or a set of participants, while specifying the type of that relation with a semantic role. Semantic annotation may focus on semantic role labeling (SRL). It labels the role of each of the participants in an event referred to by a predicate verb.

Here are two examples, one in German and another in classical Latin.

Example 1.11 German and Latin compared

a. German: Jemand hat Mia einen Ring gegeben.
b. Classical Latin:
   Arma virumque cano, Trojae qui primus ab oris
   Italiam fato profugus Lavinaque venit
   litora . . .[17]

If these sentences, especially Example 1.11a in German, are annotated with semantic roles, they can easily be translated to English.

Annotation 1.12 Semantic role labeling of the German fragment

a. Jemand$_{agent}$ hat Mia$_{recipient}$ einen Ring$_{theme}$ gegeben$_{event}$.
b. Someone$_{agent}$ has Mia$_{recipient}$ a ring$_{theme}$ given$_{event}$.
c. Translation: Someone has given Mia a ring.

[16] For example, see Shahid and Kazakov (2013) for parallel corpora with word senses related to wordNet synsets.
[17] The first three lines of Virgil's *Aeneid*.

https://doi.org/10.1017/9781108884532.004 Published online by Cambridge University Press

There are three steps to translation: (i) Annotate the source text (German) with semantic roles, (ii) translate each of the words in the source to a corresponding word with the semantic role in the target language (English), and (iii) reorder the word order in it to obtain the translation.

The identical process applies to the Latin text.

Annotation 1.13 Virgil annotated
Arma$_{theme}$ virum$_{theme}$que $\emptyset_{agent:1S}$ cano$_{event1}$,
Trojae qui primus ab oris$_{source}$
Italiam$_{goal1}$ fato profugus Lavinaque $\emptyset_{agent:3S}$ venit$_{event2}$
litora$_{goal2}$ ....[18]

We now go through the step of word-for-word translation and then reorder the words.

Translation 1.14 Virgil translated
a. Annotated translation:
warfare$_{theme}$ and a man$_{theme}$ $I_{agent:1S}$ sing$_{event1}$,
of Troy who first from the coast$_{source}$
Italy$_{goal1}$ by fate fleeing and Lavinian came$_{event2}$
shore$_{goal2}$ ....
b. Polished translation:
I sing of warfare and a man,
who, first fleeing from the coast of Troy
to Italy by fate came to the Lavinian
shore ....

Semantic role annotation is also applicable to a question-answering system (QAS). Consider a question like the following.

Example 1.15 Question annotated
a. What did Mia get from Yong?
b. What$_{theme}$ did Mia get from Yong?

To answer this question based on the annotated data, one looks for the expression which carries the semantic role of being a *theme* in that data. It should be the ring in this case.

Issues are more complicated, requiring types of semantic annotation other than semantic role labeling (SRL). Consider one more example.

[18] In Latin, every verb carries information about its Subject. Here it is represented by the emptyset symbol ∅. "1S" stands for first person singular and "3S" for third person singular.

https://doi.org/10.1017/9781108884532.004 Published online by Cambridge University Press

Annotation 1.16 Semantic role labeling
Mia$_{agent}$ left$_{pred}$ Seoul$_{initialLoc}$ for Boston$_{goal}$ yesterday$_{time}$.

This is understood to be saying that Mia was the one who departed from Seoul, she was heading for Boston, and the time of her departure was the date referred to by *yesterday*. Seoul was the location where Mia initiated her trip, while Boston was the goal or intended destination of Mia's trip.

The temporal expression *yesterday* is a so-called *indexical* expression with its reference determined contextually. The specification of the date referred to by *yesterday* depends on the utterance time, the time when the dataset was created, and also information on the time zones in Seoul and Boston. The annotation of temporal expressions requires more information than their just being labeled *time*.

Temporal annotation provides exact dates for indexical expressions like *yesterday*, although it is sometimes argued that semantic annotation should give the meaning of *yesterday*.[19]

Annotation 1.17 Temporal annotation
Mia$_{agent}$ left$_{pred}$ Seoul$_{initialLoc}$ for Boston$_{goal}$ yesterday$_{data:2018\text{-}11\text{-}01}$.

This date is calculated on the basis of the time of utterance, the time and date of data creation, and the relevant information about the time zone differences.

Temporal annotation such as TimeML can apply to the evaluation of question-answering situations like the following.

Example 1.18 Question answering

a. Q: When did Mia leave Seoul?
   A: On the first of November.
b. Q: Will she be in Boston today?
   A: She should be if she has taken a direct flight to New York.

1.4 Extended Summary

Annotation provides additional information, called *metadata*, to text or other forms of data in language. As I mentioned, it has a long scholarly tradition, especially working with ancient texts such as the Confucian Analects, the Hebrew Bible, or grammar books to explicate them.

[19] This date is not the meaning or intension of *yesterday*, but the date to which the term *yesterday* specifically (extensionally) refers.

https://doi.org/10.1017/9781108884532.004 Published online by Cambridge University Press

A variety of formats have been used to represent annotations: innotes, footnotes, sidenotes, or endnotes. The content of annotation has also varied from simple comments to detailed illustrations to supplement the main content.

Such a scholarly practice was extended to the analysis of language data, called *linguistic annotation*. First, a large amount of textual data is collected and sorted into a machine-readable set of files, called *corpus*. Second, annotation applies to such data collection, involving base segmentation, tokenization, POS-tagging, or syntactic analysis (parsing). Semantic annotation requires data segmentation as a prerequisite, while making use of morpho-syntactic analysis.

This chapter also mentioned machine learning for natural language annotation. The theory and techniques of machine learning have been adopted to train machines as well as humans to learn to work together for annotation. It has become the core of doing linguistic annotation at the current stage.

Linguistics used to be considered as consisting of three levels: phonology, morphology, and syntax. The mixing of linguistic levels was considered unscientific, especially by strict Structural Linguistics in the 1930s and 1940s. Semantics was not accepted into proper linguistics till the mid-1970s. It is now a basis for semantic annotation.

The domain of semantic annotation is much broader than that of formal semantics. Semantic annotation applies to the whole area of language processing from phonetics to pragmatics to the analysis of dialogues and discourses including the multimodal aspects of communication such as gestures and facial expressions that express a variety of sentiments. Semantic annotation works on every type of information that is relevant for communicative actions.

Semantic annotation is characterized by the partiality and situatedness of information. For example, an event-based temporal annotation scheme (e.g., TimeML) annotates those expressions, called *markables*, in a dataset that refers to time or events only. Semantic annotation provides context-specific information only. Given temporal expressions like *yesterday*, annotation specifies its exact date, not just stating that that was a day before today. Suppose someone finds a note, saying *Sorry that I had to spend a day here yesterday. Thanks,* LK. One who reads the note and wonders what date that *yesterday* refers to is not interested in knowing the meaning of the word *yesterday*. Rather than looking up a dictionary, the annotator or message breaker would look for a clue for locating the date for *yesterday* mentioned in the note.

This chapter concludes with a brief illustration of how semantic annotation can apply to some of the NLP applications. Semantic role annotation, for instance, can easily apply to machine translation (MT) and question answering (QA).