Glossary

Aboutness: a reader's perception of a text's or corpus' overall meaning or focus, separate from—but often reinforced by—the presence of keywords.
Annotation: a process of adding information to the language content of a text or corpus. Annotations may be "representational" in that they describe text (e.g., orthographic annotation, such as marking the beginning and ending of words) or "interpretive" in that they add analysis to text (e.g., pragmatic annotation, such as describing how language is used).
Balancedness: a characteristic of corpus design and creation referring to the intentional and proportional representation of all forms of language content that a corpus is meant to represent.
Collocation: a) an analytic process based on finding two or more words or phrases that occur near each other within texts in a corpus. Such words or phrases are said to be "co-located" or "collocates." b) a tool built into corpus analysis software that is capable of counting the actual and expected frequency of word or phrase co-location.
Concordance: a) a listing of every instance of a word or phrase appearing in a corpus of texts; b) a tool built into corpus analysis software that is capable of locating and listing every instance of a word or phrase appearing in a corpus of texts.
Concordance lines: a listing of every instance of a word or phrase in a corpus, including the words to the left and right which make up the word's or phrase's context.
Corpora: plural form of "corpus," referring to more than one collection of texts used in analysis. Often used when comparing two collections of text in one analysis.
Corpus: a collection of texts that share a common trait, source, subject, form, or function. The collection is usually a sample of texts that represent the larger population of texts from which they are drawn. Examples might include white papers on an emerging technology, op-eds on education, or presidential speeches.
Dictionary: a) a collection of words or phrases organized into groups that represent a shared characteristic or meaning (e.g., action verbs, hedges, negative evaluation words, evidentials); b) a resource used in computer-assisted content analysis to automate coding of that content by matching words or phrases in a corpus with a collection of other words organized in the way described in definition a.
Discourse: a) a stream of written or spoken words usually exchanged between speakers or correspondents; b) broadly, any form of intentional communication, whether between co-present participants or those who are not co-present (e.g., a writer and reader).
Dispersion: a measure of the degree to which a word or phrase is spread through

texts that comprise the corpus. A high dispersion score indicates that a word or phrase appears in many texts in a corpus. A low dispersion score indicates more sporadic spread of a word or phrase throughout a corpus. The term can also be used to measure the even or uneven spread of a word or phrase through a text or texts (e.g., terms appearing frequently at the beginning of a text but nowhere else are not evenly spread; whereas, terms appearing consistently in the beginning, middle, and end of the text are more evenly spread).

Distant Reading: an analysis method that identifies patterns describing the shape of data without examining individual data points. Examples may include word clouds, word frequency counts, and automated data coding.

Diversity: a characteristic of corpus design and creation, referring to the intentional representation of the full variety of language content (including speakers, contexts, purposes, registers, etc.) that a corpus is meant to represent.

Frequency: a) the number of times a word or phrase appears in a corpus, which may be represented as absolute (i.e., raw count) or relative (i.e., proportion of a whole) value; b) a tool built into corpus analysis software that is capable of counting occurrences of words or phrases.

Keyness: a measure of a word's importance to the meaning of a corpus or a text within that corpus. This importance may be expressed in a variety of statistical ways, including log-likelihood and log ratio. Each expression attempts to indicate whether a word's frequency is larger or smaller than what would be expected by comparison to a separate corpus used as a reference. Cf. *Positive Keyness* and *Negative Keyness*.

KWIC: a) an acronym standing for Key Word In Context, meaning a word or phrase plus its immediate textual context; b) a tool built into corpus analysis software that is capable of locating words or phrases and presenting them with words to the left or right that constitute the immediate context. Cf. *Concordance*.

Lemma: the base form of a word from which other forms may be derived to serve other grammatical and syntactic functions in discourse. For example: written, writer, writers, and writerly derive from the lemma *write*. Lemmas are often presented with an asterisk after the letters forming the lemma to indicate "all words beginning with these letters and closing with any ending" (e.g., searching for the lemma *write** would return write, writes, writer, writers, and writerly).

Lemmatization: a) the process of choosing and creating lemmas; b) the process of analyzing a collection of words derived from a single lemma.

Lexicography: an area of study within corpus linguistics referring to the analysis of word meaning, usage, and change. Lexicographic analysis may involve tracking word usage to identify changes in meaning and use over the course of time, in different contexts of use, or among different people.

N-gram: a) a unit of language describing a grouping of immediately adjacent words where "N" refers to the number of words comprising the unit (e.g., 3-gram = three-word phrase); b) a tool built into corpus analysis software that is capable of locating and listing sequential groupings of words with a specified "N" length value.

Negative Keyness: a measure of a word's lack of importance to the meaning of a corpus or a text within that corpus. This lack of importance may be expressed in a variety of statistical ways, including log-likelihood and log ratio. Each expression attempts to indicate whether a word's frequency is smaller than what would be expected by comparison to a corpus used as a reference. Cf. *Keyness* and *Positive Keyness*.

Population: the full universe of texts comprising the discourse that a corpus represents in part. For example: a corpus of software user documentation comes from a population of all software user documentation.

Positive Keyness: a measure of word's importance to the meaning of a corpus or a text within that corpus. This importance may be expressed in a variety of statistical ways, including log-likelihood and log ratio. Each expression attempts to indicate whether a word's frequency is larger than what would be expected by comparison to a corpus used as a reference. Cf. *Keyness* and *Negative Keyness*.

Proportional Representation: expresses a word's frequency in a corpus as a percentage of the whole set of words (or phrases) in the corpus. The figure may also be represented as a normalized value projected per 10,000 words. Also called "relative frequency."

Register: Uses of language that are specific to a distinctive situation. Related to the concept of genre, but more constrained to the presence or absence of recurrent words or phrases.

Relative Frequency: Cf. *proportional representation.*

Representativeness: "the extent to which a sample includes the full range of variability in a population" (Biber, 1993, p. 243).

Saturation: a point during the analysis process when a researcher stops finding examples that expand the theoretical criteria that are germane to the study. In other words, saturation is when a researcher has found all the categories or findings that are relevant to the study at hand.

Study Corpus: In comparative studies: the corpus that is under investigation. Sometimes called "target corpus." Cf. *Reference Corpus.*

Reference Corpus: In comparative studies: the corpus being used to create contrast with the study corpus. Cf. *study corpus.*

Thin Description: The process of describing a phenomena based on a limited reading of a large number of data points. The process is the opposite of thick description, which describes a phenomena based on a detailed reading of a limited number of data points. Cf. *Distant Reading.*