Glossary

AAVE: African American Vernacular English
ACE: the Australian Corpus of English, also known as the Macquarie Corpus
ACH: the Association for Computers and the Humanities
ACL: the Association for Computational Linguistics
alignment: establishing a link between the source text and the translation, usually at the sentence, phrase or word level.
ALLC: the Association for Literary and Linguistic Computing
ANC: the American National Corpus
annotation: the process of encoding interpretative linguistic information in a corpus
ARCHER: a Representative Corpus of Historical English Registers
ASCII: American Standard Code for Information Interchange
authenticity: a feature that characterizes naturally occurring corpus data
BNC: the British National Corpus
BNCweb: the web interface of the BNC, developed at Zurich University
BoE: the Bank of English
Brown: the Brown University Standard Corpus of Present-day American English
CA: contrastive analysis
CANCODE: the Cambridge and Nottingham Corpus of Discourse in English
CDA: critical discourse analysis
CED: the Corpus of English Dialects
CEPC: Chinese-English Parallel Corpus
CES: the Corpus Encoding Standard
character encoding: a system of using numeric values to represent characters
CHILDES: the Child Language Data Exchange System
chi-square test: a measure of statistical significance
CIA: Contrastive Interlanguage Analysis
CKIP: the Chinese Knowledge Information Processing group at Academia Sinica, Taipei
CLC: Cambridge Learner Corpus
CLEC: Chinese Learner English Corpus
COCOA: one of the earliest markup schemes that uses a set of attribute names and values enclosed in angled brackets
colligation: the collocation of a node word with a particular grammatical class of words
collocation: the characteristic co-occurrence of patterns of words
comparable corpus: a corpus which is composed of L1 data collected from different languages using the same sampling techniques
comparative corpus: a corpus containing components of varieties of the same language
concordance: an alphabetical index of a search pattern in a corpus, showing every contextual occurrence of the search pattern
concordancer: a software package that extracts concordances from a corpus
corpora: the widely accepted plural form of corpus
corpus balance: the range of different types of language that a corpus claims to cover
corpus header: the part of a corpus that provides necessary bibliographical information, taxonomies used and other metadata relating to a corpus

corpus: a collection of sampled texts, written or spoken, in machine readable form which may be annotated with various forms of linguistic information
corpuses: a less commonly used plural form of corpus
CPE: the Corpus of Professional English
CPSA: the Corpus of Professional Spoken American English
cross-tabulation: a table showing the frequencies for each variable across each sample
CSAE: the Corpus of South African English
DCMI: the Dublin Core Metadata Initiative
DDL: data-driven learning
dispersion: a term in descriptive statistics which refers to a quantifiable variation of measurements of differing members of a population within the scale on which they are measured
ditto tag: in corpus annotation assigning the same part-of-speech code to each word in an idiomatic expression
DTD: Document Type Definitions in markup languages such as HTML, SGML and XML
EAGLES: Expert Advisory Group on Language Engineering Standards
EAP: English for Academic Purpose
EBMT: Example-based Machine Translation
EMILLE: the Enabling Minority Language Engineering (project and corpora)
ENPC: the English-Norwegian Parallel Corpus
error-tagging: assigning codes indicating the types of errors occurring in a learner corpus
factor analysis: a statistical analysis commonly used in the social and behavioural sciences to summarize the interrelationships among a large group of variables in a concise fashion
Fisher's exact test: an alternative to the chi-square or log-likelihood test that measures exact statistical significance level
FLOB: the Freiburg-LOB Corpus of British English, an update of the LOB corpus in the early 1990s
frequency: also called raw frequency, the actual count of a linguistic feature in a corpus
Frown: the Freiburg-Brown Corpus of American English, an update of the Brown corpus in the early 1990s
HKUST: the HKUST Computer Science Corpus
HTML: Hypertext Markup Language
ICE: the International Corpus of English
ICLE: the International Corpus of Learner English
IMDI: the ISLE Metadata Initiative
IMDI: the ISLE Metadata Initiative
interlanguage: the learner's knowledge of the L2 which is independent of both the L1 and the actual L2
JEFLL: the Japanese EFL Learner Corpus
keyword: words in a corpus whose frequency is unusually high (positive keywords) or low (negative keywords) in comparison with a reference corpus
KWIC: key-word-in-context concordance
LCA: the Lancaster Corpus of Abuse
LCMC: the Lancaster Corpus of Mandarin Chinese
lemmatization: grouping together all of the different inflected forms of the same word
lexicon: an inventory of word forms in a given language

LGSWE: the Longman Grammar of Spoken and Written English
LIVAC: Linguistic Variations in Chinese Speech Communities, a synchronous corpus of Mandarin Chinese
LLC: the London-Lund Corpus; also found to refer to the Longman Learners' Corpus in the literature
LOB: the Lancaster-Oslo-Bergen Corpus of British English
LOCNESS: the Louvain Corpus of Native English Essays
log-likelihood test: also known as an LL test, an alternative to the chi-square test
LPC: the Lancaster Parsed Corpus
LSAC: the Longman Spoken American Corpus
LSP: language for specific purposes
markup: a system of standard codes inserted into a document stored in electronic form to provide information *about* the text itself and govern formatting, printing or other processing
MATE: the Multi-Level Annotation Tools Engineering project
mean: the arithmetic average, which can be calculated by adding all of the scores together and then dividing the sum by the number of scores
merger: combination of two or more words (e.g. *can't* and *gonna*)
metadata: a term used to describe data about data, typically the contextual information of corpus samples
MI: mutual information, a statistical formula borrowed from information theory
MICASE: the Michigan Corpus of Academic Spoken English
Microconcord: a concordance package published the Oxford University Press
ML: machine learning
MLCT: the Multilingual Corpus Tool package developed by Scott Songlin Piao
monitor corpus: a corpus that is constantly supplemented with fresh material and keeps increasing in size
MonoConc: a concordancer package published by Athelstan
MUC: the Message Understanding Conference
Multiconcord: a multilingual parallel concordancer developed at the University of Birmingham
MWU: multiword unit
NLP: natural language processing
normalization: a process which makes frequencies from samples of markedly different sizes comparable by bringing them to a common base
OCR: optical character recognition
OLAC: the Open Language Archives Community
ParaConc: a bilingual or multilingual concordancer published by Athelstan
parallel corpus: a corpus which is composed of source texts and their translations in one or more different languages; sometimes referred to as translation corpus
parsing: also called treebanking or bracketing, a process that analyzes the sentences in a corpus into their constituents
PERC: the Professional English Research Consortium
PNC: the Polish National Corpus
population: the entire set of items from which samples can be drawn
POS: part-of-speech
post-editing: human correction of automatically processed data
range: the difference between the highest and lowest frequencies
reference corpus: a balanced representative corpus for general usage; in keyword analysis, a corpus that is used to provide a reference wordlist

representativeness: a corpus is thought to be representative of the language variety it is supposed to represent if the findings based on its contents can be generalized to the said language variety
RP: Received Pronunciation, the notional standard form of spoken British English
sample: elements that are selected intentionally as a representation of the population being studied
sample corpus: as opposed to a monitor corpus, a sample corpus is of finite size and consists of text segments selected to provide a static snapshot of language
SARA: SGML Aware Retrieval Application for the BNC
SBCSAE: the Santa Barbara Corpus of Spoken American English
SEC: the Lancaster/IBM Spoken English Corpus
SED: the Survey of English Dialects corpus
semantic prosody: the collocational meaning arising from the interaction between a given node word and its collocates
SEU: Survey of English Usage
SGML: the Standard Generalized Markup Language
skeleton parsing: also called shallow parsing, a parsing technique that uses less fine-grained constituent types rather than would be present in a full parse
SLA: second language acquisition
sort: arrange concordances or a wordlist in a certain order
SPAAC: the Speech Act Annotated Corpus developed at UCREL, Lancaster
specialized corpus: a corpus that is domain or genre specific and is designed to represent a sublanguage
SPSS: Statistical Package for the Social Sciences
SST: the Standard Speaking Test corpus consisting of spoken data produced Japanese learners of English
standardized type-token ratio: similar to type-token ratio, but computed every *n* (e.g. 1,000) words as the WordSmith Wordlist goes through each text file
subcorpus: a component of a corpus, usually defined using certain criteria such as text types and domains
tagging: an alternative term for annotation, especially word-level annotation such as POS tagging and semantic tagging
tagset: a scheme of codes for corpus annotation, especially POS tagging
TEI: the Text Encoding Initiative
token: an occurrence of any given word form
tokenization: also called segmentation, a process that divides running text into legitimate word tokens, especially important for languages such as Chinese that do not delimit words with white spaces
transcription: converting spoken data into a written form
translationese: a version of L1 language that has been influenced by the translation process
treebank: an alternative term for a parsed corpus
t-test: an alternative statistical test to the chi-square test
type: a word form
type-token ratio: the ratio between types and tokens, useful when comparing samples of roughly equal length
UCL: University College London
UCLES: the University of Cambridge Local Examinations Syndicate
UCREL: the University Centre for Computer Corpus Research on Language, Lancaster

Unicode: a character encoding system designed to support the interchange, processing, and display of all of the written texts of the diverse languages of the world
URL: Uniform Resource Locator, i.e. an Internet address
USAS: the UCREL Semantic Analysis System
UTF: Unicode Transformation Format
wildcard: a special character such as an asterisk (*) or a question mark (?) that can be used to represent one or more characters in pattern matching
wordlist: a list of words occurring in a corpus, possibly with frequency information
WordSmith: a corpus exploration package with sophisticated statistical analysis, published by the Oxford University Press
WSC: the Wellington Corpus of Spoken New Zealand English
WWC: the Wellington Corpus of Written New Zealand English
Xaira: XML Aware Indexing and Retrieval Architecture, a new XML-aware version of SARA that can work with different corpora
Xanadu: an X-windows interactive editor for anaphoric annotation, developed at Lancaster UCREL
XCES: XML Corpus Encoding Standard
XML: the Extensible Markup Language
z-test: an alternative statistical test to chi-square test