Fall26_SLS775. Natural Language Processing in Corpus Research

Hakyung Sung, Second Language Studies, University of Hawaii at Manoa
Last updated: August 12, 2026

Contents

1 Introduction to NLP for corpus linguistics

References

- Dunn (2022). *Natural language processing for corpus linguistics*. Cambridge University Press. Chapter 1.1: Scaling Up Corpus Linguistics. [Section 1 of these notes]
- Gries (2026). Corpus linguistics and psycholinguistics. In *International encyclopedia of language and linguistics* (3rd ed., pp. 412–416). Elsevier. [Section 1.1]
- Meurers (2020), Natural language processing and language learning. *Encyclopedia of applied linguistics*, 4193-4205. [Section 1.2]

Corpus linguistics and natural language processing (NLP) both investigate language through collections of naturally occurring linguistic data. However, the two fields have traditionally differed in their primary goals. Corpus linguistics generally begins with a linguistic question and uses corpus data to describe patterns of language use, while NLP develops computational methods for automatically analyzing or generating human language. NLP for corpus linguistics brings these perspectives together: computational models are used to identify, annotate, measure, and compare linguistic patterns in corpora at a scale that would difficult to achieve through manual analysis alone.

This combination has become increasingly important as the size and diversity of available language data have expanded. Early electronic corpora, such as the Brown Corpus, contained approximately one million words, whereas contemporary corpora may contain billions of words. Such datasets create new opportunities for investigating linguistic variation, but they also make exclusively manual analysis impractical. Computational methods therefore allow corpus linguists to scale up their analyses while maintaining an explicit and, in principle, reproducible analytical procedure.

One way to organize computational problems in corpus linguistics is to distinguish between two broad types of analysis: (1) *categorization* and (2) *comparison*. Categorization involves assigning a predefined label to a linguistic unit. For example, a corpus linguist may classify a word by its part of speech, determine whether a sentence instantiates a particular grammatical construction, or identify the genre to which a document belongs. Comparison, by contrast, involves assessing the degree of similarity or relatedness between linguistic units. A researcher may examine whether two words are semantically related, whether two sentences convey similar meanings, or whether two texts share stylistic characteristics.

These two types of analysis are not rigidly tied to particular machine-learning algorithms. Note that in *supervised* learning, a model is trained on examples that have already been assigned *target labels* (in other words, annotated tags). The model learns the relationship between the input

features and the expected output and then applies that relationship to new data. In *unsupervised* learning, by contrast, the data are not provided with predefined target labels. The model instead identifies patterns, groupings, or relationships within the data.

Categorization is often implemented through supervised learning because researchers typically define the categories that a model is expected to predict. For example, a researcher may provide texts labeled as academic writing, news, or conversation and train a classifier to assign new texts to one of these registers. However, categories can also be explored through unsupervised methods. A clustering algorithm might group texts according to their linguistic similarities, after which the researcher examines whether the resulting clusters correspond to recognizable registers.

Comparison is often conducted using representations or similarity measures that do not require predefined class labels. For example, a researcher may represent two sentences as vectors and calculate their cosine similarity to estimate how closely related their meanings are. However, comparison can also involve supervised learning. A model may be trained on sentence pairs that have been labeled with degrees of semantic similarity and then used to predict the similarity of new sentence pairs.

1.1 What is corpus linguistics?

Corpus linguistics is an *empirical* approach to the study of language based on the systematic analysis of collections of linguistic data known as *corpora*. At the most basic technical level, a corpus may consist of multiple machine-readable text files accompanied by information about the texts, speakers, and contexts from which the data were collected (i.e., metadata). However, a corpus is not simply any large collection of texts. Its value as a research resource depends on how systematically the data were selected, organized, and documented. This distinction also helps clarify the difference between linguistic *data* and a *corpus*: whereas data may refer broadly to any collection of linguistic observations, a corpus is a deliberately constructed and documented dataset designed to support systematic analysis of a particular population or domain of language use.

A corpus is generally designed to represent a particular population or domain of language use, such as a group of speakers, a register, a genre, a topic, a historical period, or a language variety. Its sampling scheme should therefore reflect relevant sources of variability within the population it is intended to represent. For example, a corpus of contemporary spoken English may sample speakers from different regions, age groups, and communicative settings. Claims based on a corpus are consequently limited by the population and contexts represented in its design. We will return to this issue in our discussion of corpus *representativeness*.

Corpus linguistics is primarily concerned with the distribution of linguistic forms and functions. Researchers examine how frequently linguistic elements occur, the contexts in which they occur, the other elements with which they co-occur, and how these distributions vary across speakers, texts, registers, or time periods. These distributional patterns can be summarized through

measures (e.g., token frequency, association strength).

Corpora are commonly used in psycholinguistics as sources of quantitative measures derived from the distributional patterns of language use. These measures can serve as operationalizations of psycholinguistic constructs that are not directly obserable. In other words, corpus statistics may not be able to measure cognitive representations or processes directly, but they quantify aspects of an individual's likely linguistic experience that may be relevant to how language is (learned and) used. For example, one of the most basic corpus measures is *token frequency*, or the number of times a particular linguistic form or pattern occurs. Token frequency can be related to *entrenchment*, the process through which repeated exposure strengthens a linguistic representation and allows it to become established as a cognitive routine.[1] Similarly, the frequency with which two linguistic elements occur together may contribute to the entrenchment of that combination. Corpus data also make it possible to quantify *contingency*, or the extent to which the occurrence of one linguistic element predicts the occurrence of another. Such relationships can be investigated through analyses of collocation and/or collostruction, often using conditional probabilities or association measures. For example, researchers may examine whether a particular verb occurs with a given construction more frequently than would be expected from the independent frequencies of the verb and the construction. Strong contingency may facilitate the learning of form–form or form–meaning associations because the occurrence of one element provides reliable information about another.

One important advantage of corpus data is their potential for ecological validity.[2] Experimental studies are often conducted under highly controlled conditions using carefully selected and sometimes decontextualized stimuli. Such control is necessary for establishing causal relationships, but the resulting language may not always reflect the distributions encountered in ordinary communication. Balanced experimental designs may also expose participants to linguistic forms in proportions that differ substantially from those found in everyday language use. Corpus data complement experimental evidence by documenting how language is distributed in naturally occurring, less experimentally constrained communicative contexts.

At the same time, corpus data present several analytical challenges. Naturally occurring language is often noisy and heterogeneous, and linguistic variables are frequently highly intercorrelated. Contexts may vary considerably both within and across speakers, making it difficult to isolate the influence of any single factor. Corpus distributions are also typically highly imbalanced and approximately Zipfian: a small number of linguistic elements occur very frequently, whereas most occur only rarely. Consequently, corpus frequencies and associations must be interpreted in light of corpus composition, sampling decisions, dispersion, and the number of relevant opportunities for occurrence.

[1] For example, a word, phrase, or grammatical construction that occurs frequently in language use may be recognized or processed more readily than a less frequently encountered alternative.

[2] Ecological validity refers to "the degree to which a given research context is indicative of, or can be generalized to, real-world environments" (Schmuckler, 2026). More details here: https://oecs.mit.edu/pub/1y8kpgak

Corpus linguistics is therefore not simply the automated counting of words in large text collections. It is a research approach that connects systematically sampled language data, quantitative descriptions of linguistic distributions, and theoretically motivated interpretations of language use. Its strength lies in allowing researchers to investigate how linguistic patterns emerge across large numbers of contextualized observations while remaining attentive to the limitations of the data and the procedures used to analyze them.

1.2 What is NLP?

NLP is concerned with the automated analysis and generation of human language. NLP systems process written or spoken language in order to identify linguistic properties, extract information, make predictions, or produce new language. Common NLP tasks include dividing texts into words and sentences, assigning part-of-speech (POS) categories, identifying syntactic relationships, measuring semantic similarity, translating between languages, classifying documents, answering questions, and generating text. Speech recognition and speech synthesis are also closely related to NLP, although speech processing is sometimes treated as a separate subfield.

NLP can operate at different levels of linguistic analysis. At the word level, a system may identify lemmas, morphological features, or POS tags. At the sentence level, it may analyze grammatical structure, identify errors, or represent meaning. At the text level, it may measure cohesion, classify genre, estimate proficiency, or generate a summary. NLP therefore provides methods for transforming unstructured language into representations that can be systematically searched, counted, compared, and modeled.

This role makes NLP particularly important for corpus linguistics. A corpus may contain thousands or millions of texts that cannot be analyzed manually in their entirety. NLP tools can automatically annotate these texts and extract linguistic features relevant to a research question. For example, a researcher may use an NLP pipeline to identify verbs and their subjects, calculate lexical and syntactic complexity measures for individual texts, or represent texts as vectors for classification and comparison. NLP therefore supports the scalability of corpus analysis. However, the validity of the resulting analysis depends on several factors: the capabilities and limitations of the computational method used (e.g., a simple count-based system vs. Transformer architecture), the characteristics of the corpus to which it is applied (e.g., edited standard-language texts from news sources vs. L2 learner language), and the extent to which the computational output provides a valid operationalization of the linguistic construct under investigation (e.g., text length vs. grammatical complexity).

In the context of language learning, NLP has two broad applications. The first is the analysis of *learner language*: words, sentences, and texts produced by language learners. The second is the analysis of *language for learners*: texts and other language materials that are selected, adapted, or generated to support learning.

The first broad application includes areas such as intelligent language tutoring systems, au-

tomatic language assessment, and grammatical error detection. An intelligent language tutoring system may analyze a learner's response in order to provide individualized feedback, select an appropriate subsequent activity, or update a model of the learner's developing abilities. NLP is necessary in such systems because it is generally impossible to anticipate every response that a learner might produce and manually specify the appropriate feedback for each possible string. Instead, the system must abstract away from the exact wording of the response and identify more general properties of its form, meaning, and appropriateness. This requirement creates a distinctive challenge for NLP. Many general-purpose NLP systems are designed to remain robust when input contains errors/unexpected forms. A parser, for instance, is typically expected to produce a syntactic analysis even when a sentence is not fully well formed. In a language tutoring system, however, the unexpected form may itself be the information that the system needs to identify. The system must determine not only that a learner' s response differs from an expected target but also what kind of difference has occurred, what it may reveal about the learner' s developing language system, and what feedback would help the learner understand the problem. High correction accuracy alone is therefore not always sufficient. A system that produces an acceptable corrected sentence without identifying the linguistic source of the problem may provide limited insight into the learner' s underlying difficulty and limited support for learning.

NLP is also used to annotate and analyze learner corpora with a goal to investigate broader patterns of development and variation. NLP can support error annotation, identify overuse/underuse of linguistic forms, calculate measures of complexity, accuracy, and fluency, and examine relationships between linguistic features and proficiency. It can also help researchers investigate cross-linguistic influence by identifying patterns associated with learners' first-language backgrounds. By automating at least part of this process, NLP makes it possible to analyze larger learner corpora and connect theoretical questions more directly to relevant instances in the data. This application closely reflects the role of NLP in corpus linguistics introduced in Section 1 and will be a central focus throughout this course.

The second broad application involves analyzing language that is presented to learners. For example, NLP can help identify reading materials that contain particular vocabulary or grammatical structures, estimate the difficulty or readability of texts, highlight target forms, provide links to definitions or grammatical explanations, and generate exercises or assessment items from existing materials. Such applications can give learners more individualized access to language resources and allow instructional materials to be selected according to their proficiency, goals, or current learning needs.

We briefly discussed NLP within the context of applied linguistics, but the field can be defined more broadly when viewed from the perspectives of computer science and artificial intelligence. The following definition, adapted from Vajjala et al. (2020, p. 6), highlights this broader perspective:

NLP is an interdisciplinary field at the intersection of computer science, artificial in-

telligence, and linguistics. It focuses on developing computational systems that can process, analyze, and understand human language. Although NLP was historically concentrated in academic institutions and research laboratories, recent advances have enabled its widespread adoption in areas such as retail, healthcare, finance, law, marketing, and human resources. Several developments have contributed to this expansion:

1. the growing availability of accessible NLP tools, techniques, and application programming interfaces (APIs);
2. improvements in generalizable and interpretable approaches to complex tasks, including open-domain dialogue and question answering;
3. increased investment in interactive consumer technologies that use language as a primary means of communication;
4. the availability of open-source datasets and standardized benchmarks; and
5. the development of datasets and language-specific models for languages beyond English and other widely digitized languages.

Chapter Summary

Corpus linguistics and NLP both use collections of language data, but they have traditionally emphasized different goals. Corpus linguistics begins with a question about language use and examines the distribution of linguistic forms and functions in (systematically constructed) corpora. NLP develops computational methods for processing, analyzing, and generating language. NLP for corpus linguistics brings these perspectives together by using computational methods to annotate, search, measure, categorize, and compare linguistic patterns across datasets that may be too large for entirely manual analysis.

Activity: Brainstrom your project for this course

You might not have a concrete research question yet. Start with a topic that you are currently studying or may be interested in studying throughout the course.

1. Briefly describe your research area or a research question that interests you. What aspect of language, language use, or language learning would you like to investigate?
2. What kind of language data could help you investigate this question? For example, would you examine learner writing, classroom interaction, interviews, social media posts, published texts, assessment responses, or another type of data?
3. Could this dataset be considered a corpus? What information would you need about its speakers, writers, texts, tasks, or contexts in order to interpret it appropriately?
4. What linguistic feature, pattern, or construct would you examine in the data? Examples might include vocabulary use, grammatical constructions, errors, syntactic complexity, semantic similarity, interactional features, or differences across learner groups.
5. Would your analysis mainly involve *categorization*, *comparison*, or both? Explain your choice using one example from your research topic.
6. What part of the analysis might NLP help automate? What would you still want a researcher to examine or verify manually?
7. Identify potential challenge(s). For example, could the corpus be unrepresentative, could the NLP tool perform poorly on the language data, or could the computational measure fail to capture the intended linguistic construct?

Group report: Select one research example from your group and briefly explain (1) the research question, (2) the proposed language data, and (3) one way that NLP could support the analysis.

References

Dunn, J. (2022). *Natural language processing for corpus linguistics*. Cambridge University Press.

Gries, S. T. (2026). Corpus linguistics and psycholinguistics. In Nesi, H. and Milin, P., editors, *International Encyclopedia of Language and Linguistics*, pages 412–416. Elsevier, 3 edition.

Meurers, D. (2020). Natural language processing and language learning. *Encyclopedia of applied linguistics*, pages 4193–4205.

Schmuckler, M. A. (2026). *Ecological Validity*. MIT Press. https://oecs.mit.edu/pub/1y8kpgak.

Vajjala, S., Majumder, B., Gupta, A., and Surana, H. (2020). *Practical natural language processing: a comprehensive guide to building real-world NLP systems*. O'Reilly Media.