Natural language processing

**Natural language processing** (**NLP**) is a subfield of [computer science](https://en.wikipedia.org/wiki/Computer_science) and especially [artificial intelligence](https://en.wikipedia.org/wiki/Artificial_intelligence). It is primarily concerned with providing computers with the ability to process data encoded in [natural language](https://en.wikipedia.org/wiki/Natural_language) and is thus closely related to [information retrieval](https://en.wikipedia.org/wiki/Information_retrieval), [knowledge representation](https://en.wikipedia.org/wiki/Knowledge_representation) and [computational linguistics](https://en.wikipedia.org/wiki/Computational_linguistics), a subfield of [linguistics](https://en.wikipedia.org/wiki/Linguistics). Typically data is collected in [text corpora](https://en.wikipedia.org/wiki/Text_corpus), using either rule-based, statistical or neural-based approaches in [machine learning](https://en.wikipedia.org/wiki/Machine_learning) and [deep learning](https://en.wikipedia.org/wiki/Deep_learning).

Major tasks in natural language processing are [speech recognition](https://en.wikipedia.org/wiki/Speech_recognition), [text classification](https://en.wikipedia.org/wiki/Text_classification), [natural-language understanding](https://en.wikipedia.org/wiki/Natural-language_understanding), and [natural-language generation](https://en.wikipedia.org/wiki/Natural_language_generation).

**History**

[[edit](https://en.wikipedia.org/w/index.php?title=Natural_language_processing&action=edit&section=1)]

*Further information:*[*History of natural language processing*](https://en.wikipedia.org/wiki/History_of_natural_language_processing)

Natural language processing has its roots in the 1950s.[[1]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-1) Already in 1950, [Alan Turing](https://en.wikipedia.org/wiki/Alan_Turing) published an article titled "[Computing Machinery and Intelligence](https://en.wikipedia.org/wiki/Computing_Machinery_and_Intelligence)" which proposed what is now called the [Turing test](https://en.wikipedia.org/wiki/Turing_test) as a criterion of intelligence, though at the time that was not articulated as a problem separate from artificial intelligence. The proposed test includes a task that involves the automated interpretation and generation of natural language.

**Symbolic NLP (1950s – early 1990s)**

[[edit](https://en.wikipedia.org/w/index.php?title=Natural_language_processing&action=edit&section=2)]

The premise of symbolic NLP is well-summarized by [John Searle](https://en.wikipedia.org/wiki/John_Searle)'s [Chinese room](https://en.wikipedia.org/wiki/Chinese_room) experiment: Given a collection of rules (e.g., a Chinese phrasebook, with questions and matching answers), the computer emulates natural language understanding (or other NLP tasks) by applying those rules to the data it confronts.

* **1950s**: The [Georgetown experiment](https://en.wikipedia.org/wiki/Georgetown-IBM_experiment) in 1954 involved fully [automatic translation](https://en.wikipedia.org/wiki/Automatic_translation) of more than sixty Russian sentences into English. The authors claimed that within three or five years, machine translation would be a solved problem.[[2]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-2) However, real progress was much slower, and after the [ALPAC report](https://en.wikipedia.org/wiki/ALPAC) in 1966, which found that ten years of research had failed to fulfill the expectations, funding for machine translation was dramatically reduced. Little further research in machine translation was conducted in America (though some research continued elsewhere, such as Japan and Europe[[3]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-3)) until the late 1980s when the first [statistical machine translation](https://en.wikipedia.org/wiki/Statistical_machine_translation) systems were developed.
* **1960s**: Some notably successful natural language processing systems developed in the 1960s were [SHRDLU](https://en.wikipedia.org/wiki/SHRDLU), a natural language system working in restricted "[blocks worlds](https://en.wikipedia.org/wiki/Blocks_world)" with restricted vocabularies, and [ELIZA](https://en.wikipedia.org/wiki/ELIZA), a simulation of a [Rogerian psychotherapist](https://en.wikipedia.org/wiki/Rogerian_psychotherapy), written by [Joseph Weizenbaum](https://en.wikipedia.org/wiki/Joseph_Weizenbaum) between 1964 and 1966. Using almost no information about human thought or emotion, ELIZA sometimes provided a startlingly human-like interaction. When the "patient" exceeded the very small knowledge base, ELIZA might provide a generic response, for example, responding to "My head hurts" with "Why do you say your head hurts?". [Ross Quillian](https://en.wikipedia.org/w/index.php?title=Ross_Quillian&action=edit&redlink=1)'s successful work on natural language was demonstrated with a vocabulary of only *twenty* words, because that was all that would fit in a computer memory at the time.[[4]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-4)
* **1970s**: During the 1970s, many programmers began to write "conceptual [ontologies](https://en.wikipedia.org/wiki/Ontology_(information_science))", which structured real-world information into computer-understandable data. Examples are MARGIE (Schank, 1975), SAM (Cullingford, 1978), PAM (Wilensky, 1978), TaleSpin (Meehan, 1976), QUALM (Lehnert, 1977), Politics (Carbonell, 1979), and Plot Units (Lehnert 1981). During this time, the first [chatterbots](https://en.wikipedia.org/wiki/Chatterbots) were written (e.g., [PARRY](https://en.wikipedia.org/wiki/PARRY)).
* **1980s**: The 1980s and early 1990s mark the heyday of symbolic methods in NLP. Focus areas of the time included research on rule-based parsing (e.g., the development of [HPSG](https://en.wikipedia.org/wiki/Head-driven_phrase_structure_grammar) as a computational operationalization of [generative grammar](https://en.wikipedia.org/wiki/Generative_grammar)), morphology (e.g., two-level morphology[[5]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-5)), semantics (e.g., [Lesk algorithm](https://en.wikipedia.org/wiki/Lesk_algorithm)), reference (e.g., within Centering Theory[[6]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-6)) and other areas of natural language understanding (e.g., in the [Rhetorical Structure Theory](https://en.wikipedia.org/wiki/Rhetorical_structure_theory)). Other lines of research were continued, e.g., the development of chatterbots with [Racter](https://en.wikipedia.org/wiki/Racter) and [Jabberwacky](https://en.wikipedia.org/wiki/Jabberwacky). An important development (that eventually led to the statistical turn in the 1990s) was the rising importance of quantitative evaluation in this period.[[7]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-7)

**Statistical NLP (1990s–2010s)**

[[edit](https://en.wikipedia.org/w/index.php?title=Natural_language_processing&action=edit&section=3)]

Up until the 1980s, most natural language processing systems were based on complex sets of hand-written rules. Starting in the late 1980s, however, there was a revolution in natural language processing with the introduction of [machine learning](https://en.wikipedia.org/wiki/Machine_learning) algorithms for language processing. This was due to both the steady increase in computational power (see [Moore's law](https://en.wikipedia.org/wiki/Moore%27s_law)) and the gradual lessening of the dominance of [Chomskyan](https://en.wikipedia.org/wiki/Noam_Chomsky) theories of linguistics (e.g. [transformational grammar](https://en.wikipedia.org/wiki/Transformational_grammar)), whose theoretical underpinnings discouraged the sort of [corpus linguistics](https://en.wikipedia.org/wiki/Corpus_linguistics) that underlies the machine-learning approach to language processing.[[8]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-8)

* **1990s**: Many of the notable early successes in statistical methods in NLP occurred in the field of [machine translation](https://en.wikipedia.org/wiki/Machine_translation), due especially to work at IBM Research, such as [IBM alignment models](https://en.wikipedia.org/wiki/IBM_alignment_models). These systems were able to take advantage of existing multilingual [textual corpora](https://en.wikipedia.org/wiki/Text_corpus) that had been produced by the [Parliament of Canada](https://en.wikipedia.org/wiki/Parliament_of_Canada) and the [European Union](https://en.wikipedia.org/wiki/European_Union) as a result of laws calling for the translation of all governmental proceedings into all official languages of the corresponding systems of government. However, most other systems depended on corpora specifically developed for the tasks implemented by these systems, which was (and often continues to be) a major limitation in the success of these systems. As a result, a great deal of research has gone into methods of more effectively learning from limited amounts of data.
* **2000s**: With the growth of the web, increasing amounts of raw (unannotated) language data have become available since the mid-1990s. Research has thus increasingly focused on [unsupervised](https://en.wikipedia.org/wiki/Unsupervised_learning) and [semi-supervised learning](https://en.wikipedia.org/wiki/Semi-supervised_learning) algorithms. Such algorithms can learn from data that has not been hand-annotated with the desired answers or using a combination of annotated and non-annotated data. Generally, this task is much more difficult than [supervised learning](https://en.wikipedia.org/wiki/Supervised_learning), and typically produces less accurate results for a given amount of input data. However, there is an enormous amount of non-annotated data available (including, among other things, the entire content of the [World Wide Web](https://en.wikipedia.org/wiki/World_Wide_Web)), which can often make up for the inferior results if the algorithm used has a low enough [time complexity](https://en.wikipedia.org/wiki/Time_complexity) to be practical.

**Neural NLP (present)**

[[edit](https://en.wikipedia.org/w/index.php?title=Natural_language_processing&action=edit&section=4)]

In 2003, [word n-gram model](https://en.wikipedia.org/wiki/Word_n-gram_language_model), at the time the best statistical algorithm, was outperformed by a [multi-layer perceptron](https://en.wikipedia.org/wiki/Multi-layer_perceptron) (with a single hidden layer and context length of several words trained on up to 14 million of words with a CPU cluster in [language modelling](https://en.wikipedia.org/wiki/Language_model)) by [Yoshua Bengio](https://en.wikipedia.org/wiki/Yoshua_Bengio) with co-authors.[[9]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-9)

In 2010, [Tomáš Mikolov](https://en.wikipedia.org/wiki/Tom%C3%A1%C5%A1_Mikolov) (then a PhD student at [Brno University of Technology](https://en.wikipedia.org/wiki/Brno_University_of_Technology)) with co-authors applied a simple [recurrent neural network](https://en.wikipedia.org/wiki/Recurrent_neural_network) with a single hidden layer to language modelling,[[10]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-10) and in the following years he went on to develop [Word2vec](https://en.wikipedia.org/wiki/Word2vec). In the 2010s, [representation learning](https://en.wikipedia.org/wiki/Representation_learning) and [deep neural network](https://en.wikipedia.org/wiki/Deep_learning)-style (featuring many hidden layers) machine learning methods became widespread in natural language processing. That popularity was due partly to a flurry of results showing that such techniques[[11]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-goldberg:nnlp17-11)[[12]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-goodfellow:book16-12) can achieve state-of-the-art results in many natural language tasks, e.g., in [language modeling](https://en.wikipedia.org/wiki/Language_modeling)[[13]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-jozefowicz:lm16-13) and parsing.[[14]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-choe:emnlp16-14)[[15]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-vinyals:nips15-15) This is increasingly important [in medicine and healthcare](https://en.wikipedia.org/wiki/Artificial_intelligence_in_healthcare), where NLP helps analyze notes and text in [electronic health records](https://en.wikipedia.org/wiki/Electronic_health_record) that would otherwise be inaccessible for study when seeking to improve care[[16]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-16) or protect patient privacy.[[17]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-17)

**Approaches: Symbolic, statistical, neural networks**

[[edit](https://en.wikipedia.org/w/index.php?title=Natural_language_processing&action=edit&section=5)]

Symbolic approach, i.e., the hand-coding of a set of rules for manipulating symbols, coupled with a dictionary lookup, was historically the first approach used both by AI in general and by NLP in particular:[[18]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-winograd:shrdlu71-18)[[19]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-schank77-19) such as by writing grammars or devising heuristic rules for [stemming](https://en.wikipedia.org/wiki/Stemming).

[Machine learning](https://en.wikipedia.org/wiki/Machine_learning) approaches, which include both statistical and neural networks, on the other hand, have many advantages over the symbolic approach:

* both statistical and neural networks methods can focus more on the most common cases extracted from a corpus of texts, whereas the rule-based approach needs to provide rules for both rare cases and common ones equally.
* [language models](https://en.wikipedia.org/wiki/Language_model), produced by either statistical or neural networks methods, are more robust to both unfamiliar (e.g. containing words or structures that have not been seen before) and erroneous input (e.g. with misspelled words or words accidentally omitted) in comparison to the rule-based systems, which are also more costly to produce.
* the larger such a (probabilistic) language model is, the more accurate it becomes, in contrast to rule-based systems that can gain accuracy only by increasing the amount and complexity of the rules leading to [intractability](https://en.wikipedia.org/wiki/Intractable_problem) problems.

Although rule-based systems for manipulating symbols were still in use in 2020, they have become mostly obsolete with the advance of [LLMs](https://en.wikipedia.org/wiki/Large_language_model) in 2023.

Before that they were commonly used:

* when the amount of training data is insufficient to successfully apply machine learning methods, e.g., for the machine translation of low-resource languages such as provided by the [Apertium](https://en.wikipedia.org/wiki/Apertium) system,
* for preprocessing in NLP pipelines, e.g., [tokenization](https://en.wikipedia.org/wiki/Tokenization_(lexical_analysis)), or
* for postprocessing and transforming the output of NLP pipelines, e.g., for [knowledge extraction](https://en.wikipedia.org/wiki/Knowledge_extraction) from syntactic parses.

**Statistical approach**

[[edit](https://en.wikipedia.org/w/index.php?title=Natural_language_processing&action=edit&section=6)]

In the late 1980s and mid-1990s, the statistical approach ended a period of [AI winter](https://en.wikipedia.org/wiki/AI_winter), which was caused by the inefficiencies of the rule-based approaches.[[20]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-johnson:eacl:ilcl09-20)[[21]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-resnik:langlog11-21)

The earliest [decision trees](https://en.wikipedia.org/wiki/Decision_tree), producing systems of hard [if–then rules](https://en.wikipedia.org/wiki/Conditional_(computer_programming)#If%E2%80%93then(%E2%80%93else)), were still very similar to the old rule-based approaches. Only the introduction of hidden [Markov models](https://en.wikipedia.org/wiki/Markov_model), applied to part-of-speech tagging, announced the end of the old rule-based approach.

**Neural networks**

[[edit](https://en.wikipedia.org/w/index.php?title=Natural_language_processing&action=edit&section=7)]

*Further information:*[*Artificial neural network*](https://en.wikipedia.org/wiki/Artificial_neural_network)

A major drawback of statistical methods is that they require elaborate [feature engineering](https://en.wikipedia.org/wiki/Feature_engineering). Since 2015,[[22]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-22) the statistical approach has been replaced by the [neural networks](https://en.wikipedia.org/wiki/Artificial_neural_network) approach, using [semantic networks](https://en.wikipedia.org/wiki/Semantic_networks)[[23]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-23) and [word embeddings](https://en.wikipedia.org/wiki/Word_embedding) to capture semantic properties of words.

Intermediate tasks (e.g., part-of-speech tagging and dependency parsing) are not needed anymore.

[Neural machine translation](https://en.wikipedia.org/wiki/Neural_machine_translation), based on then-newly-invented [sequence-to-sequence](https://en.wikipedia.org/wiki/Seq2seq) transformations, made obsolete the intermediate steps, such as word alignment, previously necessary for [statistical machine translation](https://en.wikipedia.org/wiki/Statistical_machine_translation).

**Common NLP tasks**

[[edit](https://en.wikipedia.org/w/index.php?title=Natural_language_processing&action=edit&section=8)]

The following is a list of some of the most commonly researched tasks in natural language processing. Some of these tasks have direct real-world applications, while others more commonly serve as subtasks that are used to aid in solving larger tasks.

Though natural language processing tasks are closely intertwined, they can be subdivided into categories for convenience. A coarse division is given below.

**Text and speech processing**

[[edit](https://en.wikipedia.org/w/index.php?title=Natural_language_processing&action=edit&section=9)]

[**Optical character recognition**](https://en.wikipedia.org/wiki/Optical_character_recognition)**(OCR)**

Given an image representing printed text, determine the corresponding text.

[**Speech recognition**](https://en.wikipedia.org/wiki/Speech_recognition)

Given a sound clip of a person or people speaking, determine the textual representation of the speech. This is the opposite of [text to speech](https://en.wikipedia.org/wiki/Text_to_speech) and is one of the extremely difficult problems colloquially termed "[AI-complete](https://en.wikipedia.org/wiki/AI-complete)" (see above). In [natural speech](https://en.wikipedia.org/wiki/Natural_speech) there are hardly any pauses between successive words, and thus [speech segmentation](https://en.wikipedia.org/wiki/Speech_segmentation) is a necessary subtask of speech recognition (see below). In most spoken languages, the sounds representing successive letters blend into each other in a process termed [coarticulation](https://en.wikipedia.org/wiki/Coarticulation), so the conversion of the [analog signal](https://en.wikipedia.org/wiki/Analog_signal) to discrete characters can be a very difficult process. Also, given that words in the same language are spoken by people with different accents, the speech recognition software must be able to recognize the wide variety of input as being identical to each other in terms of its textual equivalent.

[**Speech segmentation**](https://en.wikipedia.org/wiki/Speech_segmentation)

Given a sound clip of a person or people speaking, separate it into words. A subtask of [speech recognition](https://en.wikipedia.org/wiki/Speech_recognition) and typically grouped with it.

[**Text-to-speech**](https://en.wikipedia.org/wiki/Text-to-speech)

Given a text, transform those units and produce a spoken representation. Text-to-speech can be used to aid the visually impaired.[[24]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-24)

[**Word segmentation**](https://en.wikipedia.org/wiki/Word_segmentation)**(**[**Tokenization**](https://en.wikipedia.org/wiki/Tokenization_(lexical_analysis))**)**

Tokenization is a process used in text analysis that divides text into individual words or word fragments. This technique results in two key components: a word index and tokenized text. The word index is a list that maps unique words to specific numerical identifiers, and the tokenized text replaces each word with its corresponding numerical token. These numerical tokens are then used in various deep learning methods.[[25]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-:0-25)

For a language like [English](https://en.wikipedia.org/wiki/English_language), this is fairly trivial, since words are usually separated by spaces. However, some written languages like [Chinese](https://en.wikipedia.org/wiki/Chinese_language), [Japanese](https://en.wikipedia.org/wiki/Japanese_language) and [Thai](https://en.wikipedia.org/wiki/Thai_language) do not mark word boundaries in such a fashion, and in those languages text segmentation is a significant task requiring knowledge of the [vocabulary](https://en.wikipedia.org/wiki/Vocabulary) and [morphology](https://en.wikipedia.org/wiki/Morphology_(linguistics)) of words in the language. Sometimes this process is also used in cases like [bag of words](https://en.wikipedia.org/wiki/Bag_of_words) (BOW) creation in data mining.[[*citation needed*](https://en.wikipedia.org/wiki/Wikipedia:Citation_needed)]

**Morphological analysis**

[[edit](https://en.wikipedia.org/w/index.php?title=Natural_language_processing&action=edit&section=10)]

[**Lemmatization**](https://en.wikipedia.org/wiki/Lemmatisation)

The task of removing inflectional endings only and to return the base dictionary form of a word which is also known as a lemma. Lemmatization is another technique for reducing words to their normalized form. But in this case, the transformation actually uses a dictionary to map words to their actual form.[[26]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-26)

[**Morphological segmentation**](https://en.wikipedia.org/wiki/Morphology_(linguistics))

Separate words into individual [morphemes](https://en.wikipedia.org/wiki/Morpheme) and identify the class of the morphemes. The difficulty of this task depends greatly on the complexity of the [morphology](https://en.wikipedia.org/wiki/Morphology_(linguistics)) (*i.e.*, the structure of words) of the language being considered. [English](https://en.wikipedia.org/wiki/English_language) has fairly simple morphology, especially [inflectional morphology](https://en.wikipedia.org/wiki/Inflectional_morphology), and thus it is often possible to ignore this task entirely and simply model all possible forms of a word (e.g., "open, opens, opened, opening") as separate words. In languages such as [Turkish](https://en.wikipedia.org/wiki/Turkish_language) or [Meitei](https://en.wikipedia.org/wiki/Meitei_language), a highly [agglutinated](https://en.wikipedia.org/wiki/Agglutination) Indian language, however, such an approach is not possible, as each dictionary entry has thousands of possible word forms.[[27]](https://en.wikipedia.org/wiki/Natural_language_processing#cite_note-27)

[**Part-of-speech tagging**](https://en.wikipedia.org/wiki/Part-of-speech_tagging)

Given a sentence, determine the [part of speech](https://en.wikipedia.org/wiki/Part_of_speech) (POS) for each word. Many words, especially common ones, can serve as multiple parts of speech. For example, "book" can be a [noun](https://en.wikipedia.org/wiki/Noun) ("the book on the table") or [verb](https://en.wikipedia.org/wiki/Verb) ("to book a flight"); "set" can be a noun, verb or [adjective](https://en.wikipedia.org/wiki/Adjective); and "out" can be any of at least five different parts of speech.

[**Stemming**](https://en.wikipedia.org/wiki/Stemming)

The process of reducing inflected (or sometimes derived) words to a base form (e.g., "close" will be the root for "closed", "closing", "close", "closer" etc.). Stemming yields similar results as lemmatization, but does so on grounds of rules, not a dictionary.