

WIKIPEDIA

# Computational linguistics

**Computational linguistics** is an interdisciplinary field concerned with the statistical or rule-based modeling of natural language from a computational perspective, as well as the study of appropriate computational approaches to linguistic questions.

Traditionally, computational linguistics was performed by computer scientists who had specialized in the application of computers to the processing of a natural language. Today, computational linguists often work as members of interdisciplinary teams, which can include regular linguists, experts in the target language, and computer scientists. In general, computational linguistics draws upon the involvement of linguists, computer scientists, experts in artificial intelligence, mathematicians, logicians, philosophers, cognitive scientists, cognitive psychologists, psycholinguists, anthropologists and neuroscientists, among others.

Computational linguistics has both theoretical and applied components. Theoretical computational linguistics focuses on issues in theoretical linguistics and cognitive science and applied computational linguistics focuses on the practical outcome of modeling human language use.<sup>[1]</sup>

The Association for Computational Linguistics defines computational linguistics as:

...the scientific study of language from a computational perspective. Computational linguists are interested in providing computational models of various kinds of linguistic phenomena.<sup>[2]</sup>

## Contents

### Origins

#### Approaches

- Developmental approaches
- Structural approaches
- Production approaches
  - Text-based interactive approach
  - Speech-based interactive approach
- Comprehension approaches

#### Applications

#### Subfields

#### Legacy

#### See also

#### References

#### Further reading

#### External links

## Origins

Computational linguistics is often grouped within the field of artificial intelligence but was present before the development of artificial intelligence. Computational linguistics originated with efforts in the United States in the 1950s to use computers to automatically translate texts from foreign languages, particularly Russian scientific journals, into English.<sup>[3]</sup> Since computers can make arithmetic (systematic) calculations much faster and more accurately than humans, it was thought to be only a short matter of time before they could also begin to process language.<sup>[4]</sup> Computational and quantitative methods are also used historically in the attempted reconstruction of earlier forms of modern languages and sub-grouping modern languages into language families. Earlier methods, such as lexicostatistics and glottochronology, have been proven to be premature and inaccurate. However, recent interdisciplinary studies that borrow concepts from biological studies, especially gene mapping, have proved to produce more sophisticated analytical tools and more reliable results.<sup>[5]</sup>

When machine translation (also known as mechanical translation) failed to yield accurate translations right away, automated processing of human languages was recognized as far more complex than had originally been assumed. Computational linguistics was born as the name of the new field of study devoted to developing algorithms and software for intelligently processing language data. The term "computational linguistics" itself was first coined by David Hays, a founding member of both the Association for Computational Linguistics (ACL) and the International Committee on Computational Linguistics (ICCL).<sup>[6]</sup>

To translate one language into another, it was observed that one had to understand the grammar of both languages, including both morphology (the grammar of word forms) and syntax (the grammar of sentence structure). To understand syntax, one had to also understand the semantics and the lexicon (or 'vocabulary'), and even something of the pragmatics of language use. Thus, what started as an effort to translate between languages evolved into an entire discipline devoted to understanding how to represent and process natural languages using computers.<sup>[7]</sup>

Nowadays research within the scope of computational linguistics is done at computational linguistics departments,<sup>[8]</sup> computational linguistics laboratories,<sup>[9]</sup> computer science departments,<sup>[10]</sup> and linguistics departments.<sup>[11][12]</sup> Some research in the field of computational linguistics aims to create working speech or text processing systems while others aim to create a system allowing human-machine interaction. Programs meant for human-machine communication are called conversational agents.<sup>[13]</sup>

## Approaches

Just as computational linguistics can be performed by experts in a variety of fields and through a wide assortment of departments, so too can the research fields broach a diverse range of topics. The following sections discuss some of the literature available across the entire field broken into four main area of discourse: developmental linguistics, structural linguistics, linguistic production, and linguistic comprehension.

### Developmental approaches

Language is a cognitive skill that develops throughout the life of an individual. This developmental process has been examined using several techniques, and a computational approach is one of them. Human language development does provide some constraints which make it harder to apply a computational method to understanding it. For instance, during language acquisition, human children are largely only exposed to positive evidence.<sup>[14]</sup> This means that during the linguistic development of an individual, the only evidence for what is a correct form is provided, and no evidence for what is not correct. This is insufficient information for a simple hypothesis testing procedure for information as complex as language,<sup>[15]</sup> and so provides certain boundaries for a computational approach to modeling language development and acquisition in an individual.

Attempts have been made to model the developmental process of language acquisition in children from a computational angle, leading to both statistical grammars and connectionist models.<sup>[16]</sup> Work in this realm has also been proposed as a method to explain the evolution of language through history. Using models, it has been shown that languages can be learned with a combination of simple input presented incrementally as the child develops better memory and longer attention span.<sup>[17]</sup> This was simultaneously posed as a reason for the long developmental period of human children.<sup>[17]</sup> Both conclusions were drawn because of the strength of the artificial neural network which the project created.

The ability of infants to develop language has also been modeled using robots<sup>[18]</sup> in order to test linguistic theories. Enabled to learn as children might, a model was created based on an affordance model in which mappings between actions, perceptions, and effects were created and linked to spoken words. Crucially, these robots were able to acquire functioning word-to-meaning mappings without needing grammatical structure, vastly simplifying the learning process and shedding light on information which furthers the current understanding of linguistic development. It is important to note that this information could only have been empirically tested using a computational approach.

As our understanding of the linguistic development of an individual within a lifetime is continually improved using neural networks and learning robotic systems, it is also important to keep in mind that languages themselves change and develop through time. Computational approaches to understanding this phenomenon have unearthed very interesting information. Using the Price Equation and Pólya urn dynamics, researchers have created a system which not only predicts future linguistic evolution but also gives insight into the evolutionary history of modern-day languages.<sup>[19]</sup> This modeling effort achieved, through computational linguistics, what would otherwise have been impossible.

It is clear that the understanding of linguistic development in humans as well as throughout evolutionary time has been fantastically improved because of advances in computational linguistics. The ability

to model and modify systems at will affords science an ethical method of testing hypotheses that would otherwise be intractable.

## Structural approaches

To create better computational models of language, an understanding of language's structure is crucial. To this end, the English language has been meticulously studied using computational approaches to better understand how the language works on a structural level. One of the most important pieces of being able to study linguistic structure is the availability of large linguistic corpora or samples. This grants computational linguists the raw data necessary to run their models and gain a better understanding of the underlying structures present in the vast amount of data which is contained in any single language. One of the most cited English linguistic corpora is the Penn Treebank.<sup>[20]</sup> Derived from widely-different sources, such as IBM computer manuals and transcribed telephone conversations, this corpus contains over 4.5 million words of American English. This corpus has been primarily annotated using part-of-speech tagging and syntactic bracketing and has yielded substantial empirical observations related to language structure.<sup>[21]</sup>

Theoretical approaches to the structure of languages have also been developed. These works allow computational linguistics to have a framework within which to work out hypotheses that will further the understanding of the language in a myriad of ways. One of the original theoretical theses on the internalization of grammar and structure of language proposed two types of models.<sup>[15]</sup> In these models, rules or patterns learned increase in strength with the frequency of their encounter.<sup>[15]</sup> The work also created a question for computational linguists to answer: how does an infant learn a specific and non-normal grammar (Chomsky Normal Form) without learning an overgeneralized version and getting stuck?<sup>[15]</sup> Theoretical efforts like these set the direction for research to go early in the lifetime of a field of study, and are crucial to the growth of the field.

Structural information about languages allows for the discovery and implementation of similarity recognition between pairs of text utterances.<sup>[22]</sup> For instance, it has recently been proven that based on the structural information present in patterns of human discourse, conceptual recurrence plots can be used to model and visualize trends in data and create reliable measures of similarity between natural textual utterances.<sup>[22]</sup> This technique is a strong tool for further probing the structure of human discourse. Without the computational approach to this question, the vastly complex information present in discourse data would have remained inaccessible to scientists.

Information regarding the structural data of a language is available for English as well as other languages, such as Japanese.<sup>[23]</sup> Using computational methods, Japanese sentence corpora were analyzed and a pattern of log-normality was found in relation to sentence length.<sup>[23]</sup> Though the exact cause of this lognormality remains unknown, it is precisely this sort of information which computational linguistics is designed to uncover. This information could lead to further important discoveries regarding the underlying structure of Japanese and could have any number of effects on the understanding of Japanese as a language. Computational linguistics allows for very exciting additions to the scientific knowledge base to happen quickly and with very little room for doubt.

In recent days, the structural data of languages is available for several languages of the world other than English language. Computational linguistics work is in progress on Sindhi language because the structure, grammar and domain of Sindhi language is different from the other languages of the World. The computational linguistics models for English language are not suitable for Sindhi language. Viewing this, computational linguistics work on Sindhi language <sup>[24][25][26]</sup> has been started properly by developing methods, algorithms, linguistics tools (<https://sindhilp.com/>), machine learning models and deep learning models since 2016 <sup>[27][28][29][30][31][32]</sup> to focus and solve the linguistics problems of Sindhi language. This work could lead to further important discoveries regarding the underlying structure of Sindhi, and could have any number of effects on the understanding of Sindhi as a language. Computational linguistics allows for very exciting additions to the scientific knowledge base to happen quickly and with very little room for doubt.

Without a computational approach to the structure of linguistic data, much of the information that is available now would still be hidden under the vastness of data within any single language. Computational linguistics allows scientists to parse huge amounts of data reliably and efficiently, creating the possibility for discoveries unlike any seen in most other approaches.

## Production approaches

The production of language is equally as complex in the information it provides and the necessary skills which a fluent producer must have. That is to say, comprehension is only half the problem of communication. The other half is how a system produces language, and computational linguistics has made interesting discoveries in this area.

In a now famous paper published in 1950 Alan Turing proposed the possibility that machines might one day have the ability to "think". As a thought experiment for what might define the concept of thought in machines, he proposed an "imitation test" in which a human subject has two text-only conversations, one with a fellow human and another with a machine attempting to respond like a human. Turing proposes that if the subject cannot tell the difference between the human and the machine, it may be concluded that the machine is capable of thought.<sup>[33]</sup> Today this test is known as the Turing test and it remains an influential idea in the area of artificial intelligence.

One of the earliest and best-known examples of a computer program designed to converse naturally with humans is the ELIZA program developed by Joseph Weizenbaum at MIT in 1966. The program emulated a Rogerian psychotherapist when responding to written statements and questions posed by a user. It appeared capable of understanding what was said to it and responding intelligently, but in truth, it simply followed a pattern matching routine that relied on only understanding a few keywords in each sentence. Its responses were generated by recombining the unknown parts of the sentence around properly translated versions of the known words. For example, in the phrase "It seems that you hate me" ELIZA understands "you" and "me" which matches the general pattern "you [some words] me", allowing ELIZA to update the words "you" and "me" to "I" and "you" and replying "What makes you think I hate you?". In this example ELIZA has no understanding of the word "hate", but it is not required for a logical response in the context of this type of psychotherapy.<sup>[34]</sup>

Some projects are still trying to solve the problem which first started computational linguistics off as its field in the first place. However, methods have become more refined, and consequently, the results generated by computational linguists have become more enlightening. To improve computer translation, several models have been compared, including hidden Markov models, smoothing techniques, and the specific refinements of those to apply them to verb translation.<sup>[35]</sup> The model which was found to produce the most natural translations of German and French words was a refined alignment model with a first-order dependence and a fertility model. They also provide efficient training algorithms for the models presented, which can give other scientists the ability to improve further on their results. This type of work is specific to computational linguistics and has applications that could vastly improve understanding of how language is produced and comprehended by computers.

Work has also been done in making computers produce language in a more naturalistic manner. Using linguistic input from humans, algorithms have been constructed which are able to modify a system's style of production based on a factor such as linguistic input from a human, or more abstract factors like politeness or any of the five main dimensions of personality.<sup>[36]</sup> This work takes a computational approach via parameter estimation models to categorize the vast array of linguistic styles we see across individuals and simplify it for a computer to work in the same way, making human-computer interaction much more natural.

### Text-based interactive approach

Many of the earliest and simplest models of human-computer interaction, such as ELIZA for example, involve a text-based input from the user to generate a response from the computer. By this method, words typed by a user trigger the computer to recognize specific patterns and reply accordingly, through a process known as keyword spotting.

### Speech-based interactive approach

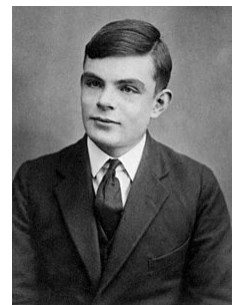
Recent technologies have placed more of an emphasis on speech-based interactive systems. These systems, such as Siri of the iOS operating system, operate on a similar pattern-recognizing technique as that of text-based systems, but with the former, the user input is conducted through speech recognition. This branch of linguistics involves the processing of the user's speech as sound waves and the interpreting of the acoustics and language patterns for the computer to recognize the input.<sup>[37]</sup>

## Comprehension approaches

Much of the focus of modern computational linguistics is on comprehension. With the proliferation of the internet and the abundance of easily accessible written human language, the ability to create a program capable of understanding human language would have many broad and exciting possibilities, including improved search engines, automated customer service, and online education.

Early work in comprehension included applying Bayesian statistics to the task of optical character recognition, as illustrated by Bledsoe and Browning in 1959 in which a large dictionary of possible letters was generated by "learning" from example letters and then the probability that any one of those learned examples matched the new input was combined to make a final decision.<sup>[38]</sup> Other attempts at applying Bayesian statistics to language analysis included the work of Mosteller and Wallace (1963) in which an analysis of the words used in *The Federalist Papers* was used to attempt to determine their authorship (concluding that Madison most likely authored the majority of the papers).<sup>[39]</sup>

In 1971 Terry Winograd developed an early natural language processing engine capable of interpreting naturally written commands within a simple rule-governed environment. The primary language parsing program in this project was called SHRDLU, which was capable of carrying out a somewhat natural conversation with the user giving it commands, but only within the scope of the toy environment designed for the task. This environment consisted of different shaped and colored blocks, and SHRDLU was capable of interpreting commands such as "Find a block which is taller than the one you are holding and put it into the box." and asking questions such as "I don't understand which pyramid you mean." in response to the user's input.<sup>[40]</sup> While impressive, this kind of natural language processing has proven much more difficult outside the limited scope of the toy environment. Similarly, a project developed by NASA called LUNAR was designed to provide answers to naturally written questions about the geological analysis of lunar rocks returned by the Apollo missions.<sup>[41]</sup> These kinds of problems are referred to as question answering.



Alan Turing: computer scientist and namesake developer of the Turing test as a method of measuring the intelligence of a machine.



Joseph Weizenbaum: former MIT professor and computer scientist who developed ELIZA, a primitive computer program utilizing natural language processing.

Initial attempts at understanding spoken language were based on work done in the 1960s and 1970s in signal modeling where an unknown signal is analyzed to look for patterns and to make predictions based on its history. An initial and somewhat successful approach to applying this kind of signal modeling to language was achieved with the use of hidden Markov models as detailed by Rabiner in 1989.<sup>[42]</sup> This approach attempts to determine probabilities for the arbitrary number of models that could be being used in generating speech as well as modeling the probabilities for various words generated from each of these possible models. Similar approaches were employed in early speech recognition attempts starting in the late 70s at IBM using word/part-of-speech pair probabilities.<sup>[43]</sup>

More recently these kinds of statistical approaches have been applied to more difficult tasks such as topic identification using Bayesian parameter estimation to infer topic probabilities in text documents.<sup>[44]</sup>

## Applications

Modern computational linguistics is often a combination of studies in computer science and programming, math, particularly statistics, language structures, and natural language processing. Combined, these fields most often lead to the development of systems that can recognize speech and perform some task based on that speech. Examples include speech recognition software, such as Apple's Siri feature, spellcheck tools, speech synthesis programs, which are often used to demonstrate pronunciation or help the disabled, and machine translation programs and websites, such as Google Translate.<sup>[45]</sup>

Computational linguistics can be especially helpful in situations involving social media and the Internet. For example, filters in chatrooms or on website searches require computational linguistics. Chat operators often use filters to identify certain words or phrases and deem them inappropriate so that users cannot submit them.<sup>[45]</sup> Another example of using filters is on websites. Schools use filters so that websites with certain keywords are blocked from children to view. There are also many programs in which parents use Parental controls to put content filters in place. Computational linguists can also develop programs that group and organize content through Social media mining. An example of this is Twitter, in which programs can group tweets by subject or keywords.<sup>[46]</sup> Computational linguistics is also used for document retrieval and clustering. When doing an online search, documents and websites are retrieved based on the frequency of unique labels related to what was typed into a search engine. For instance, if a person searches "red, large, four-wheeled vehicle," to find pictures of a red truck, the search engine will still find the information desired by matching words such as "four-wheeled" with "car".<sup>[47]</sup>

Computational approaches are also important in the study of change over time. This occurs in several ways: through the computational modeling of language families<sup>[48]</sup> (see further Quantitative comparative linguistics or phylogenetics), through the modeling of changes in sound<sup>[49]</sup> and meaning<sup>[50]</sup>

## Subfields

Computational linguistics can be divided into major areas depending upon the medium of the language being processed, whether spoken or textual; and upon the task being performed, whether analyzing language (recognition) or synthesizing language (generation).

Speech recognition and speech synthesis deal with how spoken language can be understood or created using computers. Parsing and generation are sub-divisions of computational linguistics dealing respectively with taking language apart and putting it together. Machine translation remains the sub-division of computational linguistics dealing with having computers translate between languages. The possibility of automatic language translation, however, has yet to be realized and remains a notoriously hard branch of computational linguistics.<sup>[51]</sup>

Some of the areas of research that are studied by computational linguistics include:

- Computational complexity of natural language, largely modeled on automata theory, with the application of context-sensitive grammar and linearly bounded Turing machines.
- Computational semantics comprises defining suitable logics for linguistic meaning representation, automatically constructing them and reasoning with them
- Computer-aided corpus linguistics, which has been used since the 1970s as a way to make detailed advances in the field of discourse analysis<sup>[52]</sup>
- Design of parsers or chunkers for natural languages
- Design of taggers like POS-taggers (part-of-speech taggers)
- Machine translation as one of the earliest and most difficult applications of computational linguistics draws on many subfields.
- Simulation and study of language evolution in historical linguistics/glottochronology.

## Legacy

The subject of computational linguistics has had a recurring impact on popular culture:

- The 1983 film *WarGames* features a young computer hacker who interacts with an artificially intelligent supercomputer.<sup>[53]</sup>
- A 1997 film, *Conceiving Ada*, focuses on Ada Lovelace, considered one of the first computer scientists, as well as themes of computational linguistics.<sup>[54]</sup>
- *Her*, a 2013 film, depicts a man's interactions with the "world's first artificially intelligent operating system."<sup>[55]</sup>
- The 2014 film *The Imitation Game* follows the life of computer scientist Alan Turing, developer of the Turing Test.<sup>[56]</sup>
- The 2015 film *Ex Machina* centers around human interaction with artificial intelligence.<sup>[57]</sup>
- The 2016 film *Arrival (film)*, based on Ted Chiang's *Story of Your Life*, takes a whole new approach of linguistics to communicate with advanced alien race called heptapods.<sup>[58]</sup>

## See also

- Artificial intelligence in fiction
- Collostructional analysis
- Computational lexicology
- *Computational Linguistics* (journal)
- Computational models of language acquisition
- Computational semantics
- Computational semiotics
- Computer-assisted reviewing
- Dialog systems
- Glottochronology
- Grammar induction
- Human speechome project
- Internet linguistics
- Lexicostatistics
- Natural language processing
- Natural language user interface
- Quantitative linguistics
- Semantic relatedness
- Semantometrics
- Systemic functional linguistics
- Translation memory
- Universal Networking Language

## References

- Uszkoreit, Hans. "What Is Computational Linguistics?" ([http://www.coli.uni-saarland.de/~hansu/w hat\\_is\\_cl.html](http://www.coli.uni-saarland.de/~hansu/w hat_is_cl.html)). Department of Computational Linguistics and Phonetics of Saarland University.
- "What Is Computational Linguistics?" (<http://www.aclweb.org/archive/misc/what.html>). The Association for Computational Linguistics. February 2005.
- John Hutchins: Retrospect and prospect in computer-based translation. (<http://www.hutchinsweb.me.uk/MTS-1999.pdf>) Proceedings of MT Summit VII, 1999, pp. 30–44.
- Arnold B. Barach: Translating Machine (<https://www.flickr.com/photos/bostworld/2152048032/in/set-72157603898383698/>) 1975: And the Changes To Come.
- T. Crowley., C. Bower. An Introduction to Historical Linguistics. Auckland, N.Z.: Oxford UP, 1992. Print.
- "Deceased members" (<http://nlp.shef.ac.uk/iccl/committee.html#deceased>). *ICCL members*. Retrieved 15 November 2017.
- Natural Language Processing by Liz Liddy, Eduard Hovy, Jimmy Lin, John Prager, Dragomir Radev, Lucy Vanderwende, Ralph Weischedel (<http://www-nlpir.nist.gov/MINDS/FINAL/NLP.web.pdf>)
- "Computational Linguistics and Phonetics" (<http://www.coli.uni-saarland.de>).
- "Yatsko's Computational Linguistics Laboratory" (<http://yatsko.zohosites.com/home.html>).
- "CLIP" ([https://wiki.umiacs.umd.edu/clip/index.php/Main\\_Page](https://wiki.umiacs.umd.edu/clip/index.php/Main_Page)).
- Computational Linguistics – Department of Linguistics – Georgetown College (<http://linguistics.georgetown.edu/programs/graduate/computational/>)
- "UPenn Linguistics: Computational Linguistics" (<http://www.ling.upenn.edu/research/computational.html>).
- Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N.J.: Pearson Prentice Hall.
- Bowerman, M. (1988). The "no negative evidence" problem: How do children avoid constructing an overly general grammar. Explaining language universals ([http://pubman.mpdl.mpg.de/pubman/item/escidoc:468143:4/component/escidoc:532427/bowerman\\_1988\\_The-No.pdf](http://pubman.mpdl.mpg.de/pubman/item/escidoc:468143:4/component/escidoc:532427/bowerman_1988_The-No.pdf)).
- Braine, M.D.S. (1971). On two types of models of the internalization of grammars. In D.I. Slobin (Ed.), *The ontogenesis of grammar: A theoretical perspective*. New York: Academic Press.
- Powers, D.M.W. & Turk, C.C.R. (1989). *Machine Learning of Natural Language*. Springer-Verlag. ISBN 978-0-387-19557-5.
- Elman, Jeffrey L. (1993). "Learning and development in neural networks: The importance of starting small". *Cognition*. **48** (1): 71–99. doi:10.1016/0010-0277(93)90058-4 (<https://doi.org/10.1016%2F0010-0277%2893%2990058-4>). PMID 8403835 (<https://pubmed.ncbi.nlm.nih.gov/8403835/>).
- Salvi, G.; Montesano, L.; Bernardino, A.; Santos-Victor, J. (2012). "Language bootstrapping: learning word meanings from the perception-action association". *IEEE transactions on systems, man, and cybernetics. Part B, Cybernetics: a publication of the IEEE Systems, Man, and Cybernetics Society*. **42** (3): 660–71. doi:10.1109/TSMCB.2011.2172420 (<https://doi.org/10.1109%2FTSMCB.2011.2172420>).

19. Gong, T.; Shuai, L.; Tamariz, M. & Jäger, G. (2012). E. Scalas (ed.). "Studying Language Change Using Price Equation and Pólya-urn Dynamics" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3299756>). *PLOS ONE*. **7** (3): e33171. Bibcode:2012PLoS...733171G (<https://ui.adsabs.harvard.edu/abs/2012PLoS...733171G>). doi:10.1371/journal.pone.0033171 (<https://doi.org/10.1371%2Fjournal.pone.0033171>). PMC 3299756 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3299756>). PMID 22427981 (<https://pubmed.ncbi.nlm.nih.gov/22427981>).
20. Marcus, M. & Marcinkiewicz, M. (1993). "Building a large annotated corpus of English: The Penn Treebank" (<https://www.aclweb.org/anthology/J/J93/J93-2004.pdf>) (*Computational Linguistics*. **19** (2): 313–330).
21. Taylor, Ann (2003). "1". *Treebanks*. Spring Netherlands. pp. 5–22.
22. Angus, D.; Smith, A. & Wiles, J. (2012). "Conceptual recurrence plots: revealing patterns in human discourse" (<https://espace.library.uq.edu.au/view/UQ:269425/ERAUQ269425.pdf>) (PDF). *IEEE Transactions on Visualization and Computer Graphics*. **18** (6): 988–97. doi:10.1109/TVCG.2011.100 (<https://doi.org/10.1109%2FTVCG.2011.100>). PMID 22499664 (<http://pubmed.ncbi.nlm.nih.gov/22499664>).
23. Furuhashi, S. & Hayakawa, Y. (2012). "Lognormality of the Distribution of Japanese Sentence Lengths". *Journal of the Physical Society of Japan*. **81** (3): 034004. Bibcode:2012JPSJ...81c4004F (<https://ui.adsabs.harvard.edu/abs/2012JPSJ...81c4004F>). doi:10.1143/JPSJ.81.034004 (<https://doi.org/10.1143%2FJPSJ.81.034004>).
24. "Mazhar Ali Dootio | PhD (Computer Science) Continue from SZABIST Karachi Sindh Pakistan | Independent Researcher | Computer Science | ResearchGate" ([https://www.researchgate.net/profile/Mazhar\\_Dootio](https://www.researchgate.net/profile/Mazhar_Dootio)). *ResearchGate*. Retrieved 2019-07-16.
25. "Mazhar Ali Dootio - Google Scholar Citations" (<https://scholar.google.com.pk/citations?user=s0L9vIkAAA&hl=en>). *scholar.google.com.pk*. Retrieved 2019-07-16.
26. "Sindhi NLP" (<https://sindhinlp.com/contact.php>). *sindhinlp.com*. Retrieved 2019-07-16.
27. Dootio, Mazhar Ali; Wagan, Asim Imdad (February 2019). "Development of Sindhi text corpus". *Journal of King Saud University - Computer and Information Sciences*. doi:10.1016/j.jksuci.2019.02.002 (<https://doi.org/10.1016%2Fj.jksuci.2019.02.002>). ISSN 1319-1578 (<https://www.worldcat.org/issn/1319-1578>).
28. Dootio, Mazhar Ali; Wagan, Asim Imdad (January 2019). "Syntactic parsing and supervised analysis of Sindhi text". *Journal of King Saud University - Computer and Information Sciences*. **31** (1): 105–112. doi:10.1016/j.jksuci.2017.10.004 (<https://doi.org/10.1016%2Fj.jksuci.2017.10.004>). ISSN 1319-1578 (<https://www.worldcat.org/issn/1319-1578>).
29. Wagan, Asim Imdad; Ali, Mazhar (2019-01-01). "An Analysis of Sindhi Annotated Corpus using Supervised Machine Learning Methods" (<http://publications.muett.edu.pk/index.php/muetri/article/view/754>). *Mehran University Research Journal of Engineering and Technology*. **38** (1): 185–196. Bibcode:2019MURJE...38..185A (<https://ui.adsabs.harvard.edu/abs/2019MURJE...38..185A>). doi:10.22581/muet1982.1901.15 (<https://doi.org/10.22581%2Fmuet1982.1901.15>). ISSN 2413-7219 (<https://www.worldcat.org/issn/2413-7219>).
30. Dootio, Mazhar Ali (2017). "AUTOMATIC STEMMING AND LEMMATIZATION PROCESS FOR SINDHI TEXT" (<https://www.researchgate.net/publication/328202210>), *Computational Linguistics and Intelligent Text Processing*, JSSIR NED University of Engineering and Technology Karachi Sindh Pakistan, **6**, pp. 103–112
31. Dootio, Mazhar Ali; Wagan, Asim Imdad (August 2018). "Unicode-8 based linguistics data set of annotated Sindhi text" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6139473>). *Data in Brief*. **19**: 1504–1514. doi:10.1016/j.dib.2018.05.062 (<https://doi.org/10.1016%2Fdib.2018.05.062>). ISSN 2352-3409 (<https://www.worldcat.org/issn/2352-3409>). PMC 6139473 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6139473>). PMID 30225294 (<https://pubmed.ncbi.nlm.nih.gov/30225294>).
32. "An analysis and solution of computational linguistics problems of Sindhi text" (<https://www.researchgate.net/publication/330855516>). *ResearchGate*. Retrieved 2019-07-16.
33. Turing, A. M. (1950). "Computing machinery and intelligence". *Mind*. **59** (236): 433–460. doi:10.1093/mind/lix.236.433 (<https://doi.org/10.1093%2Fmind/lix.236.433>). JSTOR 2251299 (<https://www.jstor.org/stable/2251299>).
34. Weizenbaum, J. (1966). "ELIZA—a computer program for the study of natural language communication between man and machine". *Communications of the ACM*. **9** (1): 36–45. doi:10.1145/365153.365168 (<https://doi.org/10.1145%2F365153.365168>).
35. Och, F. J.; Ney, H. (2003). "A Systematic Comparison of Various Statistical Alignment Models". *Computational Linguistics*. **29** (1): 19–51. doi:10.1162/089120103321337421 (<https://doi.org/10.1162%2F089120103321337421>).
36. Mairesse, F. (2011). "Controlling user perceptions of linguistic style: Trainable generation of personality traits". *Computational Linguistics*. **37** (3): 455–488. doi:10.1162/COLI\_a\_00063 ([http://s://doi.org/10.1162%2FCOLI\\_a\\_00063](http://s://doi.org/10.1162%2FCOLI_a_00063)).
37. *Language Files*. The Ohio State University Department of Linguistics. 2011. pp. 624–634. ISBN 9780814251799.
38. Bledsoe, W. W. & Browning, I. (1959). *Pattern recognition and reading by machine*. Papers presented at the December 1–3, 1959, eastern joint IRE-AIEE-ACM computer conference on – IRE-AIEE-ACM '59 (Eastern). New York, New York, USA: ACM Press. pp. 225–232. doi:10.1145/1460299.1460326 (<https://doi.org/10.1145%2F1460299.1460326>).
39. Mosteller, F. (1963). "Inference in an authorship problem". *Journal of the American Statistical Association*. **58** (302): 275–309. doi:10.2307/2283270 (<https://doi.org/10.2307%2F2283270>). JSTOR 2283270 (<https://www.jstor.org/stable/2283270>).
40. Winograd, T. (1971). "Procedures as a Representation for Data in a Computer Program for Understanding Natural Language" (<http://www.dtic.mil/docs/citations/AD0721399>) (Report).
41. Woods, W.; Kaplan, R. & Nash-Webber, B. (1972). "The lunar sciences natural language information system" (<https://www.researchgate.net/publication/247926251>) (Report).
42. Rabiner, L. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition". *Proceedings of the IEEE*. **77** (2): 257–286. CiteSeerX 10.1.1.381.3454 (<https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.381.3454>). doi:10.1109/5.18626 (<https://doi.org/10.1109%2F5.18626>).
43. Bahl, L.; Baker, J.; Cohen, P.; Jelinek, F. (1978). "Recognition of continuously read natural corpus". *Acoustics, Speech, and Signal*. **3**: 422–424. doi:10.1109/ICASSP.1978.1170402 (<https://doi.org/10.1109%2FICASSP.1978.1170402>).
44. Blei, D. & Ng, A. (2003). "Latent dirichlet allocation" (<http://dl.acm.org/citation.cfm?id=944937>). *The Journal of Machine Learning*. **3**: 993–1022.
45. "Careers in Computational Linguistics" (<http://www.cla.csulb.edu/departments/linguistics/careers-computational-linguistics/>). California State University. Retrieved 19 September 2016.
46. Marujo, Luís et al. "Automatic Keyword Extraction on Twitter." Language Technologies Institute, Carnegie Mellon University, n.d. Web. 19 Sept. 2016.
47. "Computational Linguistics" (<https://plato.stanford.edu/entries/computational-linguistics/#DocRetCiuApp>). *Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Feb 26, 2014. Retrieved Apr 19, 2017.
48. Bowers, Claire. "Computational phylogenetics." Annual Review of Linguistics 4 (2018): 281-296.
49. Pigoli, Davide, et al. "The analysis of acoustic phonetic data: exploring differences in the spoken romance languages." arXiv preprint arXiv:1507.07587 985 (2015); Group, The Functional Phylogenies. "Phylogenetic inference for function-valued traits: speech sound evolution." Trends in ecology & evolution 27.3 (2012): 160-166.
50. e.g. Hamilton, William L., Jure Leskovec, and Dan Jurafsky. "Diachronic word embeddings reveal statistical laws of semantic change." arXiv preprint arXiv:1605.09096 (2016).
51. Oettinger, A. G. (1965). *Computational Linguistics* (<https://www.jstor.org/stable/2313322>). The American Mathematical Monthly, Vol. 72, No. 2, Part 2: Computers and Computing, pp. 147–150.
52. McEney, Thomas (1996). *Corpus Linguistics: An Introduction* ([https://books.google.com/?id=nwmgdVn\\_akAC&pg=PA114&lpg=PA114&dq=%22computer+aided+corpus+linguistics%22#v=onepage&q=%22computer%20aided%20corpus%20linguistics%22&f=false](https://books.google.com/?id=nwmgdVn_akAC&pg=PA114&lpg=PA114&dq=%22computer+aided+corpus+linguistics%22#v=onepage&q=%22computer%20aided%20corpus%20linguistics%22&f=false)). Edinburgh: Edinburgh University Press. p. 114. ISBN 978-0748611652.
53. Badham, John (1983-06-03), *WarGames* (<https://www.imdb.com/title/tt0086567/>), retrieved 2016-02-22
54. Hershman-Leeson, Lynn (1999-02-19), *Conceiving Ada* (<https://www.imdb.com/title/tt0118882/>), retrieved 2016-02-22
55. Jonze, Spike (2014-01-10), *Her* (<https://www.imdb.com/title/tt1798709/>), retrieved 2016-02-18
56. Tyldum, Morten (2014-12-25), *The Imitation Game* ([https://www.imdb.com/title/tt2084970/?ref\\_n\\_v\\_sr\\_1](https://www.imdb.com/title/tt2084970/?ref_n_v_sr_1)), retrieved 2016-02-18
57. Garland, Alex (2015-04-24), *Ex Machina* (<https://www.imdb.com/title/tt0470752/>), retrieved 2016-02-18
58. Villeneuve, Denis (2016-10-10), "Arrival" (<https://www.imdb.com/title/tt2543164/>). Retrieved 18 December 2019.

## Further reading

- Bates, M (1995). "Models of natural language understanding" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC40721>). *Proceedings of the National Academy of Sciences of the United States of America*. **92** (22): 9977–9982. Bibcode:1995PNAS...92.9977B (<https://ui.adsabs.harvard.edu/abs/1995PNAS...92.9977B>). doi:10.1073/pnas.92.22.9977 (<https://doi.org/10.1073%2Fpnas.92.22.9977>). PMC 40721 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC40721>). PMID 7479812 (<https://pubmed.ncbi.nlm.nih.gov/7479812>).
- Steven Bird, Ewan Klein, and Edward Loper (2009). *Natural Language Processing with Python*. O'Reilly Media. ISBN 978-0-596-51649-9.
- Daniel Jurafsky and James H. Martin (2008). *Speech and Language Processing*, 2nd edition. Pearson Prentice Hall. ISBN 978-0-13-187321-6.
- Mohamed Zakaria KURDI (2016). *Natural Language Processing and Computational Linguistics: speech, morphology, and syntax*, Volume 1. ISTE-Wiley. ISBN 978-1848218482.
- Mohamed Zakaria KURDI (2017). *Natural Language Processing and Computational Linguistics: semantics, discourse, and applications*, Volume 2. ISTE-Wiley. ISBN 978-1848219212.

## External links

- Association for Computational Linguistics (ACL) (<http://www.aclweb.org/>)
  - ACL Anthology of research papers (<http://www.aclweb.org/anthology>)
  - ACL Wiki for Computational Linguistics (<http://aclweb.org/aclwiki/>)
- CILing annual conferences on Computational Linguistics (<http://www.CiLing.org/>)
- Computational Linguistics – Applications workshop (<https://web.archive.org/web/20110122142133/http://www.cla.imcsit.org/>)
- Free online introductory book on Computational Linguistics (<https://web.archive.org/web/20080125103030/http://www.gelbukh.com/clbook/>) at the Wayback Machine (archived January 25, 2008)
- Language Technology World (<https://web.archive.org/web/20180212202132/http://www.lt-world.org/>)
- Resources for Text, Speech and Language Processing (<https://www.cs.technion.ac.il/~gabr/resources/resources.html>)
- The Research Group in Computational Linguistics (<http://clg.wlv.ac.uk/>)

Retrieved from "https://en.wikipedia.org/w/index.php?title=Computational\_Linguistics&oldid=961037570"

This page was last edited on 6 June 2020, at 08:43 (UTC).

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.