

Computational Linguistics (CL) & Corpus Linguistics

Lecture 1

Outline

- ▶ The notion of **Computational Linguistics**. The **object** of its investigation. The relation with other linguistic sciences.
- ▶ The notion of **Corpus Linguistics**. The **object** of its investigation. The relation with other linguistic sciences.
- ▶ Computational vs Corpus Linguistics.



Computational Linguistics

- ▶ **Computational linguistics (CL)** is the application of computer science to the analysis, synthesis and comprehension of written and spoken language [Bernstein, 2018].
- ▶ **Computational linguistics** is the scientific and engineering discipline concerned with understanding written and spoken language from a computational perspective, and building artifacts that usefully process and produce language, either in bulk or in a dialogue setting [[Stanford Encyclopedia of Philosophy](#)].



Computational Linguistics

- ▶ “Human knowledge is expressed in language. *So computational linguistics is very important.*”

Mark Steedman, ACL Presidential Address (2007)



Vale Martin Kay: What is computational linguistics?

- ▶ **Computational linguistics** is the scientific study of language from a computational perspective. Computational linguists are interested in providing computational models of various kinds of linguistic phenomena. These models may be "**knowledge-based**" ("hand-crafted") or "**data-driven**" ("statistical" or "empirical"). Work in computational linguistics is in some cases motivated from a scientific perspective in that one is trying to provide a computational explanation for a particular linguistic or psycholinguistic phenomenon; and in other cases the motivation may be more purely technological in that one wants to provide a working component of a speech or natural language system. Indeed, the work of computational linguists is incorporated into many working systems today, including speech recognition systems, text-to-speech synthesizers, automated voice response systems, web search engines, text editors, language instruction materials, to name just a few.

Computational Linguistics is needed for

instant machine translation

- an **automatic translation** from one language to another. It makes possible to translate large passages of text in a very short time.

speech recognition (**SR**) systems

- is the translation of spoken words into text. It is also known as “automatic speech recognition”, “ASR”, “computer speech recognition”, “speech to text”, or just “STT”.

text-to-speech (TTS) synthesizers

- is the technology which lets computer speak to you. It is an application that converts text into spoken word, by analyzing and processing the text using **NLP** (Natural Language Processing) and then using **DSP** (Digital Signal Processing) technology to convert a processed text into synthesized speech representation of the text.

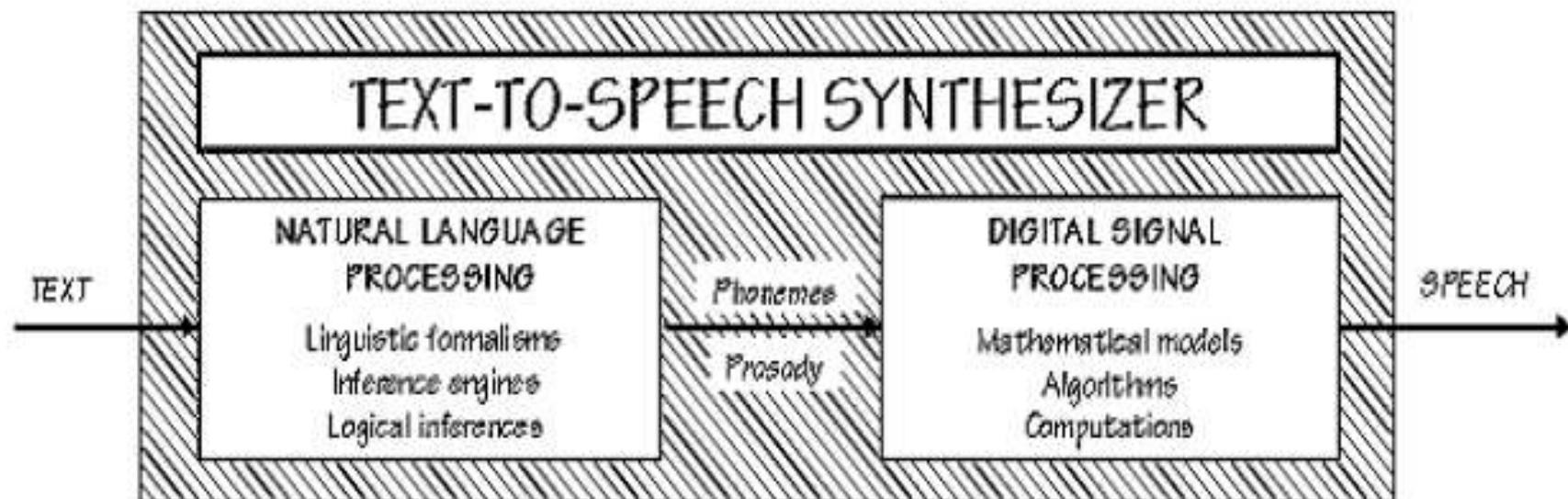


Figure 1: A simple but general functional diagram of a TTS system. [2]

Computational Linguistics is needed for

**interactive
voice response
(IVR) systems**

- an automated business phone system feature that interacts with callers and gathers information by giving them choices via a menu.

search engines

- a software system that is designed to carry out web searches.

text editors

- an app that allows you to create, open, and edit text files on your computer and Google Drive.
-



Goals of Computational Linguistics

Translating text from one language to another.

Retrieving text that relates to a specific topic.

Analyzing text or spoken language for context, sentiment or other affective qualities.

Answering questions, including those that require inference and descriptive or discursive answers.

Summarizing text.

Creating chatbots.

Computational Linguistics

requires knowledge of

machine learning
(ML)

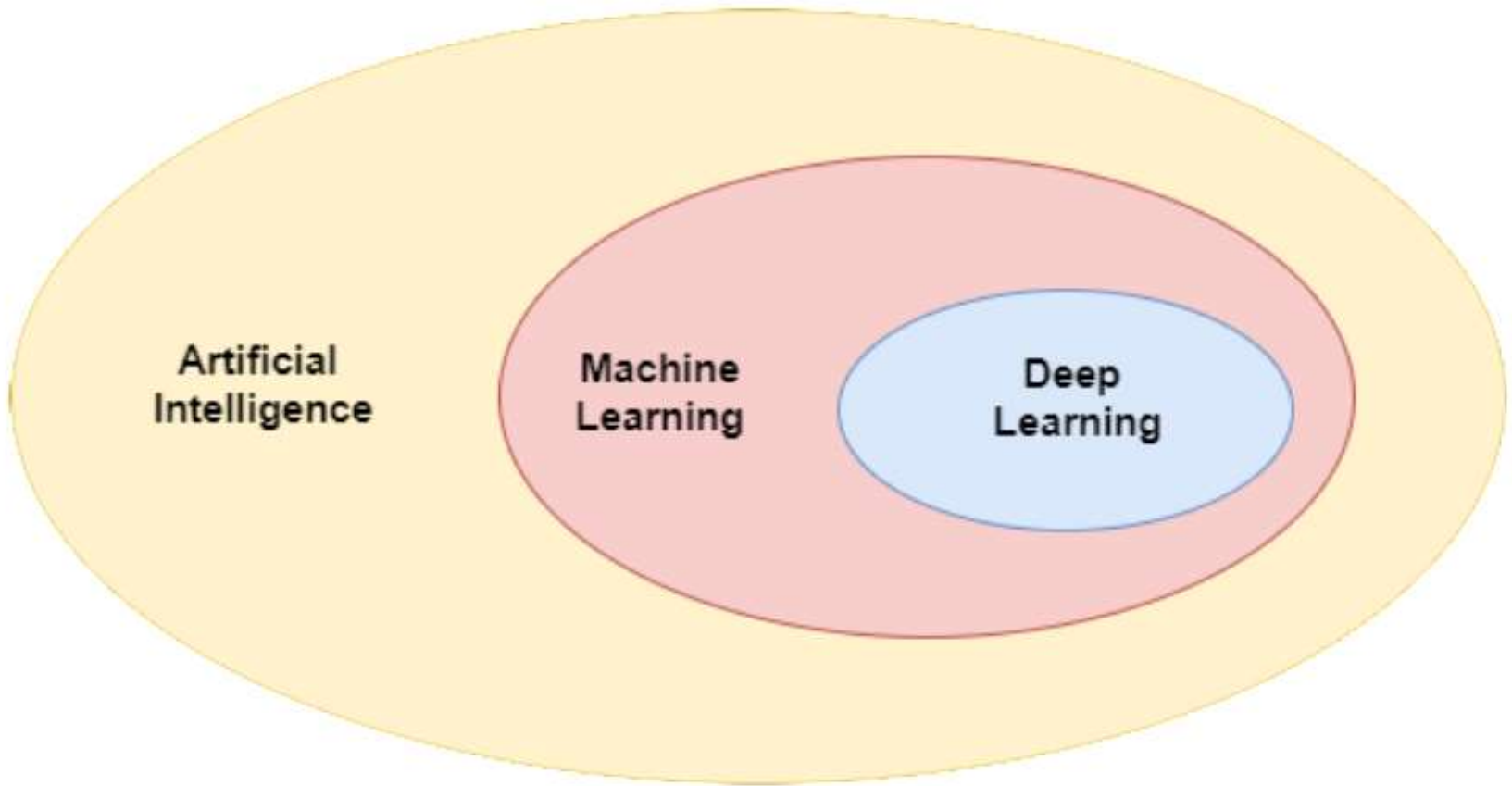
- **a method of data analysis that automates analytical model building.** It is a branch of artificial intelligence based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention [SAS].

deep learning
(DL)

- is a machine learning technique that is inspired by the way a human brain filters information, it is basically learning from examples. It helps a computer model to filter the input data through layers to predict and classify information.

artificial
intelligence (AI)

- the ability of a computer or computer-controlled robot to perform tasks commonly associated with intelligent beings.



Computational Linguistics

requires knowledge of

cognitive
computing

- refers to the use of reasoning, language processing, machine learning, and human capabilities that help regular computing better solve problems and analyze data. It enables the computer system to tackle complex decision-making processes.

neuroscience

- is a multidisciplinary science that is concerned with the study of the structure and function of the nervous system.
-

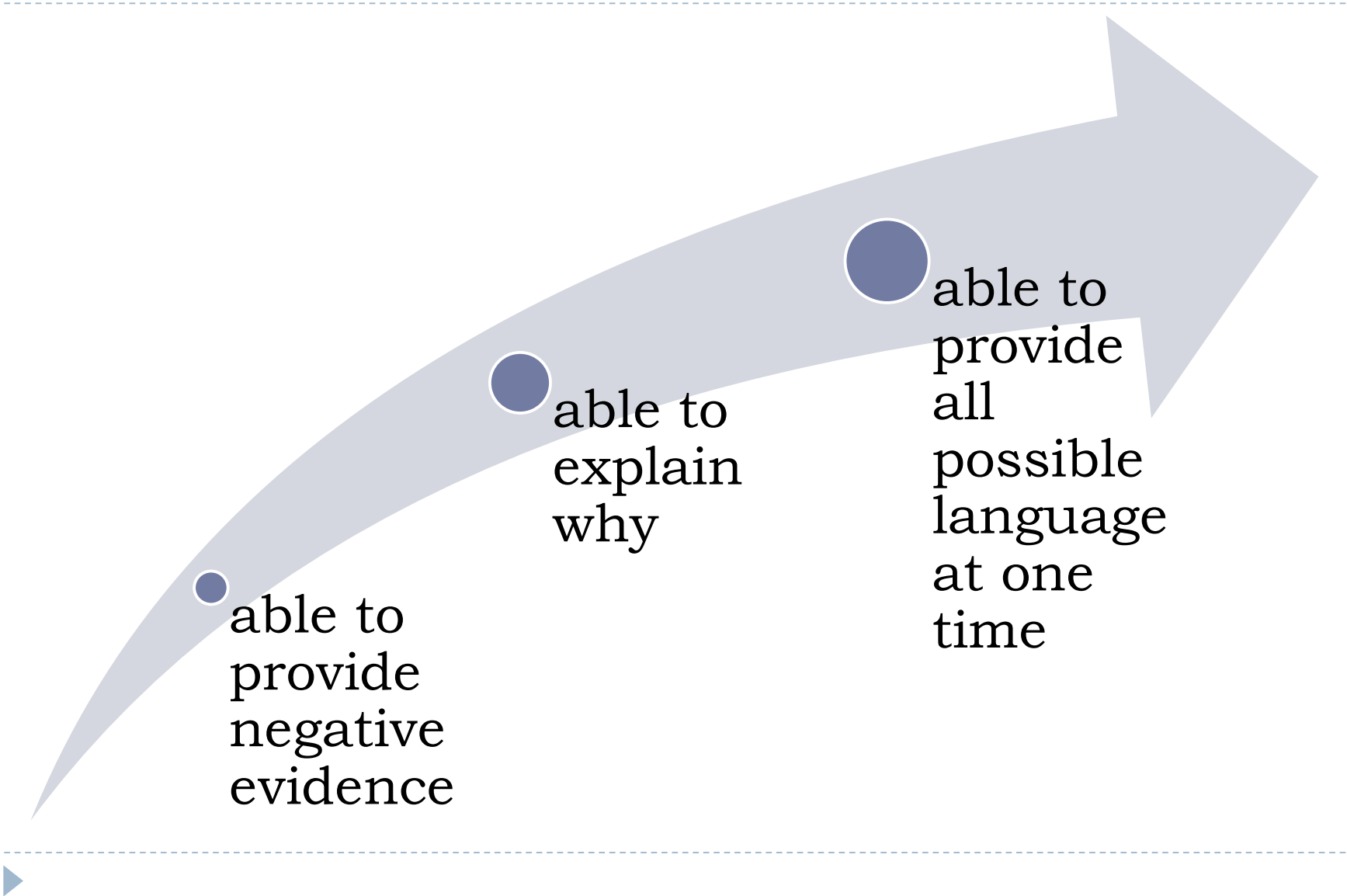


Corpus Linguistics

- ▶ **Corpus linguistics** is the study of language based on large collections of "real life" language use stored in **corpora** (or corpuses) – computerized databases created for linguistic research. It is also known as **corpus-based studies** [[Nordquist](#), 2019].



Corpus Linguistics is not



able to
provide
negative
evidence

able to
explain
why

able to
provide
all
possible
language
at one
time

Corpus Linguistics finds out

What are the most frequent words and phrases?

What are the differences between written and spoken language?

What tenses do people use most frequently?

What prepositions follow particular verbs?

How often do people use idiomatic expressions?



Corpus Approach

(Biber, Conrad&Reppen, 1998)

It is empirical, analyzing the actual patterns of language use in natural texts.

It utilizes a large and principled collection of natural texts as the basis for analysis.

It makes extensive use of computers for analysis.

It depends on both quantitative and qualitative analytical techniques.



References:

- ▶ Баранов А. Н. Введение в прикладную лингвистику: Учебное пособие. М.: Эдиториал УРСП, 2001.
 - Волошин В.Г. Комп'ютерна лінгвістика: Навч. посібник. Суми: Університетська книга, 2004.
 - Дарчук Н. П. Комп'ютерна лінгвістика (автоматичне опрацювання тексту): підручник. К.: Видавничо-поліграфічний центр "Київський університет", 2008.
 - Карпіловська Є. А. Вступ до комп'ютерної лінгвістики. Донецьк: Юго-Восток, 2003.
 - Лук'янчук С. Комп'ютерна модель парадигматичних класів дієслів // Українське мовознавство. 2000. Вип. 22. С. 82-85.
 - Пещак М. М. Нариси з комп'ютерної лінгвістики. Ужгород: Видавництво закарпаття, 1999. 200 с.
 - Bolshakov, Igor. A., Gelbukh Alexander Computational Linguistics. Models, Resources, Applications. México, 2004.
 - The Oxford Handbook of Computation Linguistics / ed. by Ruslan Mitkov. Oxford Un-ty Press, 2003.
-

Why do I need to study Computational and Corpus Linguistics?

