


Introduction to language and natural language processing (NLP)



Jasper Kyle Catapang

AI Engineer - Silicon Valley, USA

MA Applied Linguistics Student - University of Birmingham

August 19, 2021

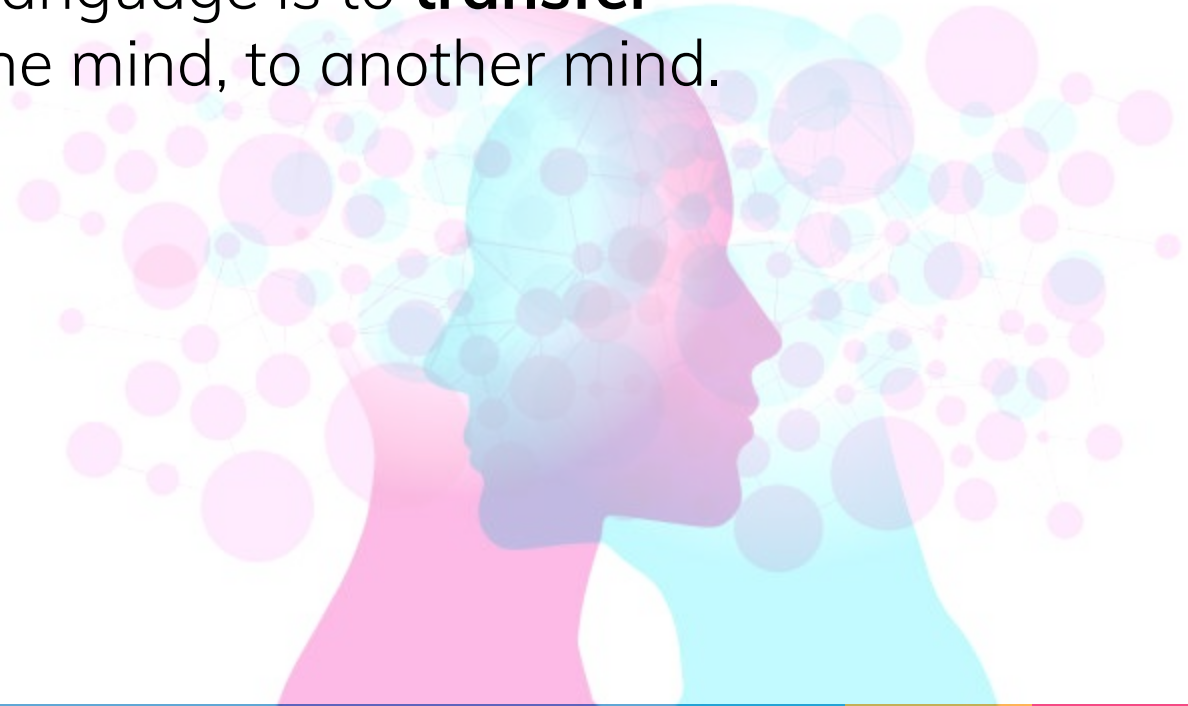
What is language?

“the principal method of human **communication**, consisting of **words** used in a **structured** and conventional way and conveyed by speech, writing, or gesture”

- Oxford Languages

What is language?

The main use of language is to **transfer thoughts** from one mind, to another mind.



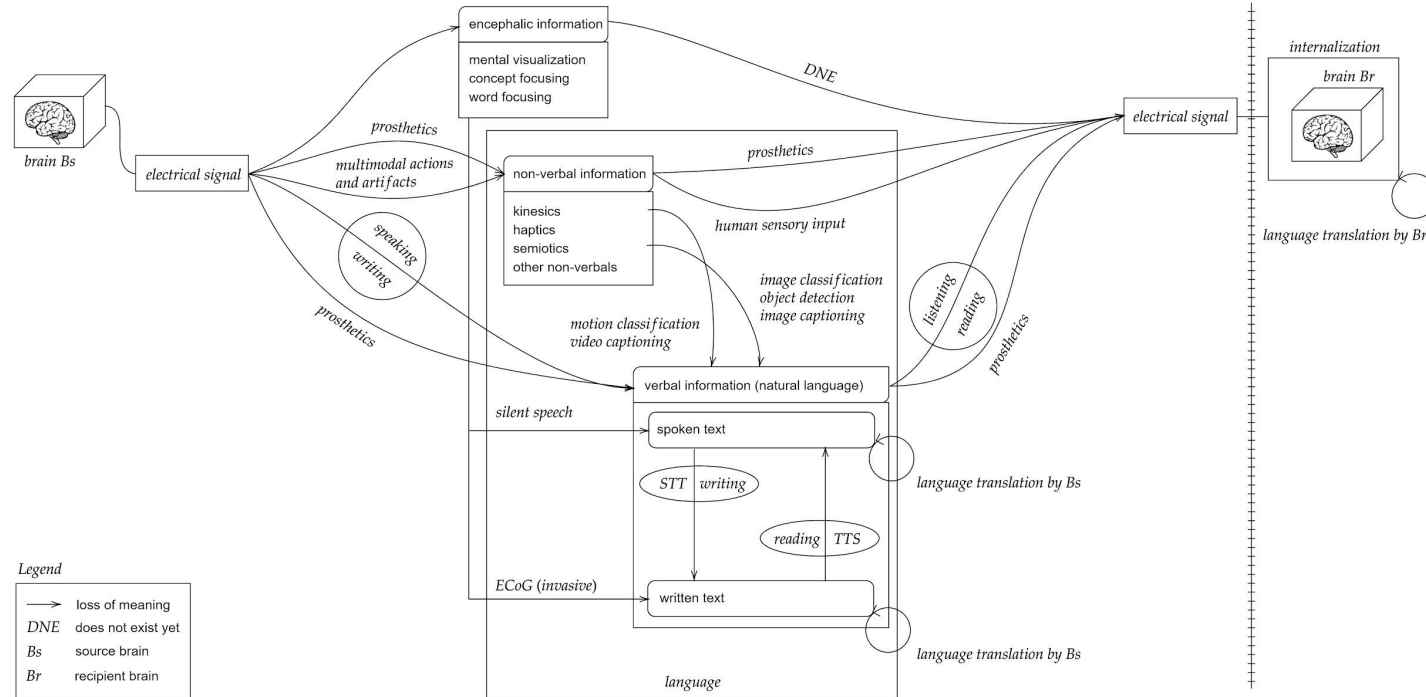
What is language?

Thought comes first, while language is an **expression**. There are certain **limitations** among language, and humans **cannot express all** that they think (Gleitman, 2005).

Propagation of meaning

Meaning is therefore **lost** with every medium or channel it goes through, from a source's mind to a recipient's mind.

Propagation of meaning model



Non-verbal to verbal example



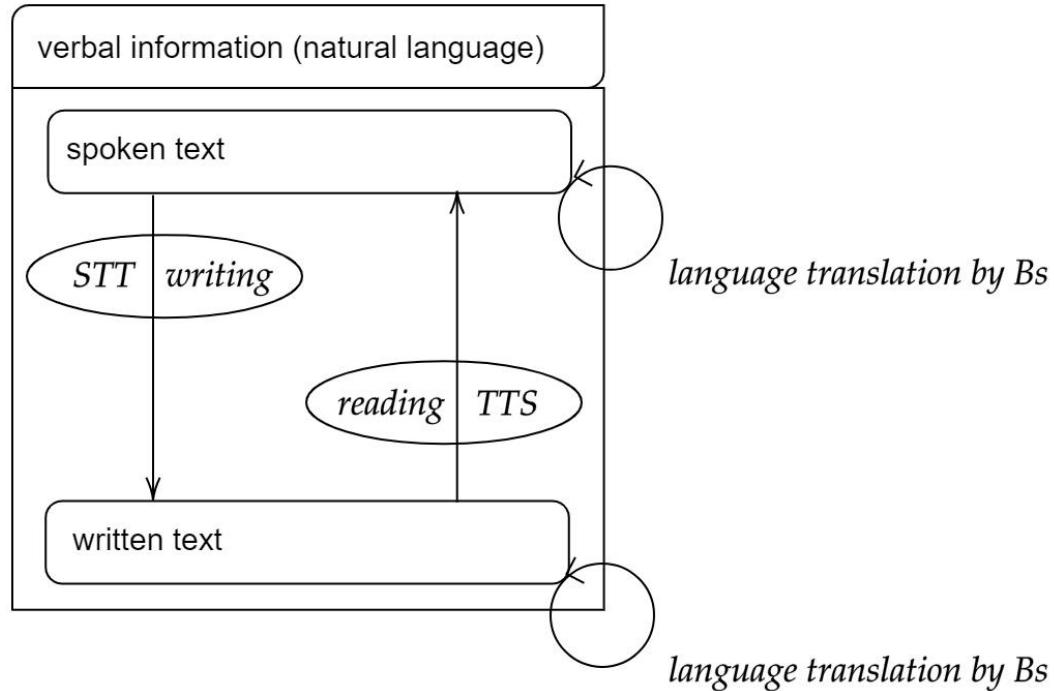
The K-pop girl group Twice is dancing to their new song.

A picture is worth a thousand words



<https://openai.com/blog/dall-e/>

Natural language



Speech to text



Google Cloud

Speech to Text



Convert your speech to text right now

Select a language and click "Start Now" to begin recording

Language

English (United States) ▼

Punctuation



Input type

☒ Microphone ☐ File upload

[Show JSON ▼](#)

 **START NOW**

28/02/2017

Optical character recognition

From Wikipedia, the free encyclopedia

Optical character recognition (also **optical character reader**) is the conversion of images of typed, handwritten or printed text in a document, a photo of a document, a scene-photo (like a photograph of a photo) or from subtitle text superimposed on an image (for example from a video recording) into machine-readable text. It can be used as a form of information entry from printed paper data records, such as bank statements, computerised receipts, business cards, mail, printouts of documents, or any other form of documentation. It is a common method of digitising printed texts so that they can be electronically searched, stored more compactly, displayed on-line, and used in machine processes such as optical character recognition, machine translation, (extracted) text-to-speech, key data and text mining. OCR is a part of pattern recognition, artificial intelligence and computer vision.

Early versions needed to be trained with images of each character, and worked on one font at a time. Modern systems capable of producing a high degree of recognition accuracy for most fonts are now available. Some systems are capable of recognising support for a variety of digital image file format inputs.^[2] Some systems are capable of recognising input that closely approximates the original page including images, columns, and other graphical elements.

28/02/2017

Optical character recognition

Optical character recognition

Optical character recognition (also **optical character reading**) is the automated conversion of images of typed, handwritten or printed text into machine-readable text. A document, a photo of a document, a scene-photo (for example from a surveillance camera) or from subtitle text superimposed on an image (for example from a television broadcast) can be used as a form of information entry from printed paper data records, which can be electronic bank statements, computerised receipts, business cards, mail, printouts of spreadsheets, or any other documentation. It is a common method of digitising printed texts so that they can be electronically searched, stored more compactly, displayed on-line, and used in machine processes such as automatic computing, machine translation, (extracted) text-to-speech, key data and text mining. OCR is a field in pattern recognition, artificial intelligence and computer vision.

Systems needed to be trained with images of each character, and worked on one font at a time, producing a high degree of recognition accuracy for most fonts are now common. Some systems are capable of recognizing image file format inputs. [2] Some systems are capable of recognizing a scanned page including images, columns, and other elements.

Early versions needed to be trained with images of each character, and worked on one font at a time. Systems capable of producing a high degree of recognition accuracy for most fonts are now common. Some systems are capable of representing that closely approximates the original page including images, columns, and other

Lost in translation

English: We bring you this show through the cooperation of our sponsors

Japanese: この番組はご覧のスポンサーの提供でお送りします

Korean: 이 프로그램은 보시는 스폰서 제공으로 보내드립니다

Filipino: Ang programang ito ay ipinadala ng sponsor na nakikita mo

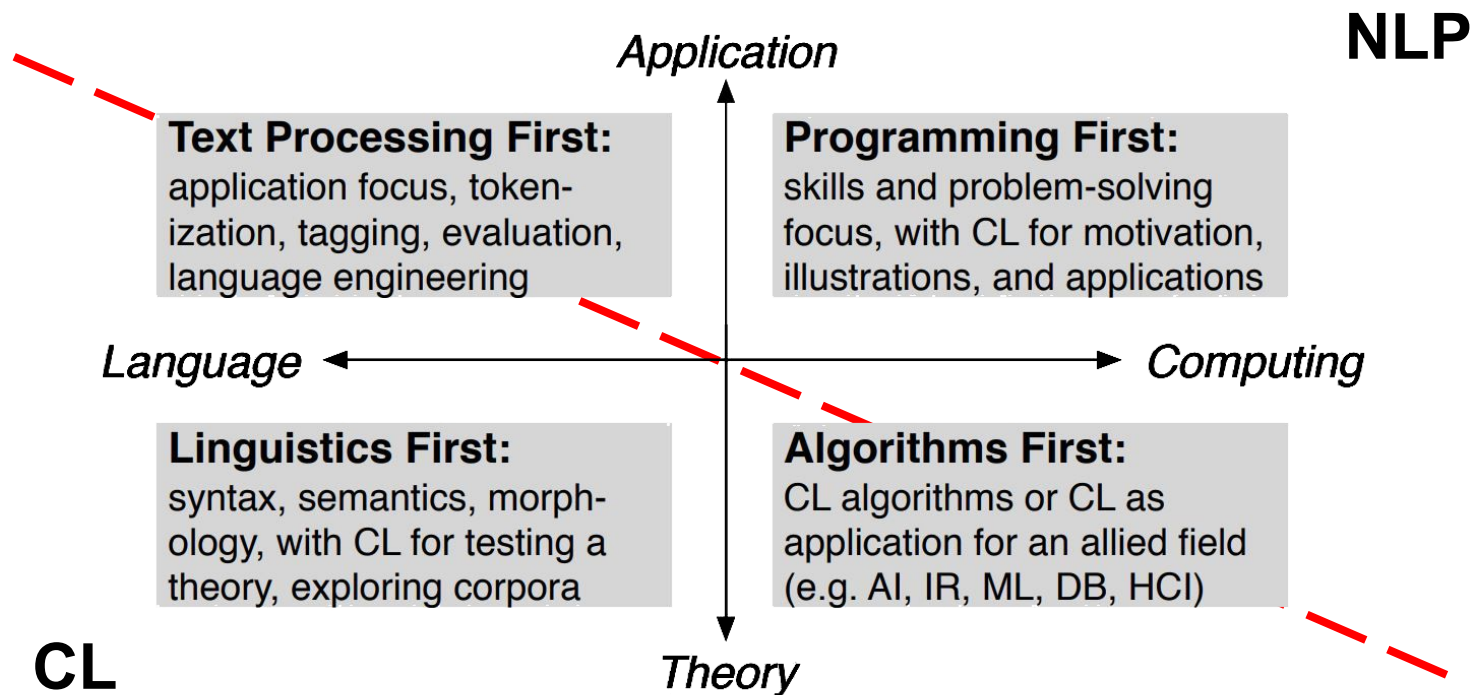
Natural language processing (NLP)

The goal of natural language processing is to get computers to perform useful tasks involving human language, tasks like **enabling human-machine** communication, **improving human-human** communication, or simply doing useful processing of text or speech (Jurafsky and Martin, 2009).

Additional goal of NLP

Keep the loss of meaning at a minimum to maximize communication.

NLP vs. computational linguistics (CL)



Why learn NLP?

- ▷ Many applications!
- ▷ Major contributor to growing AI market
- ▷ Demand is high
- ▷ Understanding how machines can learn language more efficiently and correctly

Why learn NLP?

- ▷ Many applications!
- ▷ Major contributor to growing AI market
- ▷ Demand is high
- ▷ Understanding how machines can learn language more efficiently and correctly

Applications

- ▷ Named entity recognition
- ▷ Machine translation
- ▷ Question answering
- ▷ Chatbots
- ▷ Summarization

And many more!

Named entity recognition

When **Sebastian Thrun** PERSON started at **Google** ORG in **2007** DATE , few people outside of the company took him seriously. "I can tell you very senior CEOs of major **American** NORP car companies would shake my hand and turn away because I wasn't worth talking to," said **Thrun** PERSON , now the co-founder and CEO of online higher education startup Udacity, in an interview with **Recode** ORG **earlier this week** DATE .

A little **less than a decade later** DATE , dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

Applications

- ▷ Named entity recognition
- ▷ Machine translation
- ▷ Question answering
- ▷ Chatbots
- ▷ Summarization

And many more!

Why learn NLP?

- ▷ Many applications!
- ▷ Major contributor to growing AI market
- ▷ Demand is high
- ▷ Understanding how machines can learn language more efficiently and correctly

Growing AI market

The global artificial intelligence (AI) software market is forecast to grow rapidly in the coming years, reaching around **126 billion** U.S. dollars by 2025. The overall AI market includes a wide array of applications such as natural language processing, robotic process automation, and machine learning (Liu, 2020).

Why learn NLP?

- ▷ Many applications!
- ▷ Major contributor to growing AI market
- ▷ Demand is high
- ▷ Understanding how machines can learn language more efficiently and correctly

Demand is high

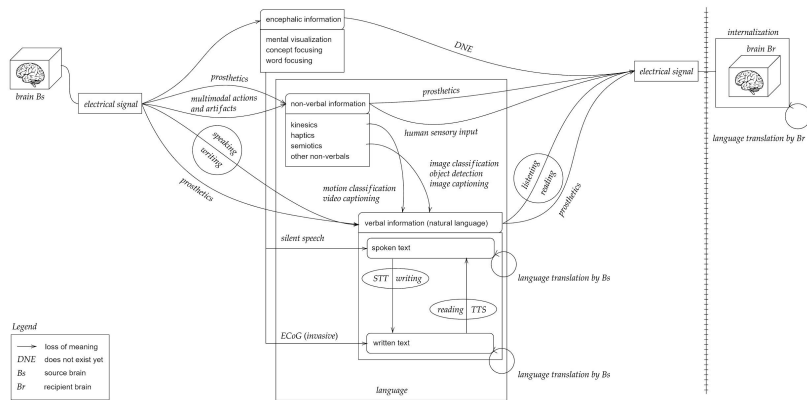
- ▷ The pandemic catalyzed digitalization of all companies and businesses
- ▷ This includes the digitalization and processing of text data
- ▷ Jobstreet openings with “natural language processing” as query, pay PHP 50,000 at the least (as of August 12, 2021).

Why learn NLP?

- ▷ Many applications!
- ▷ Major contributor to growing AI market
- ▷ Demand is high
- ▷ Understanding how machines can learn language more efficiently and correctly
 - and by extension... customers also.

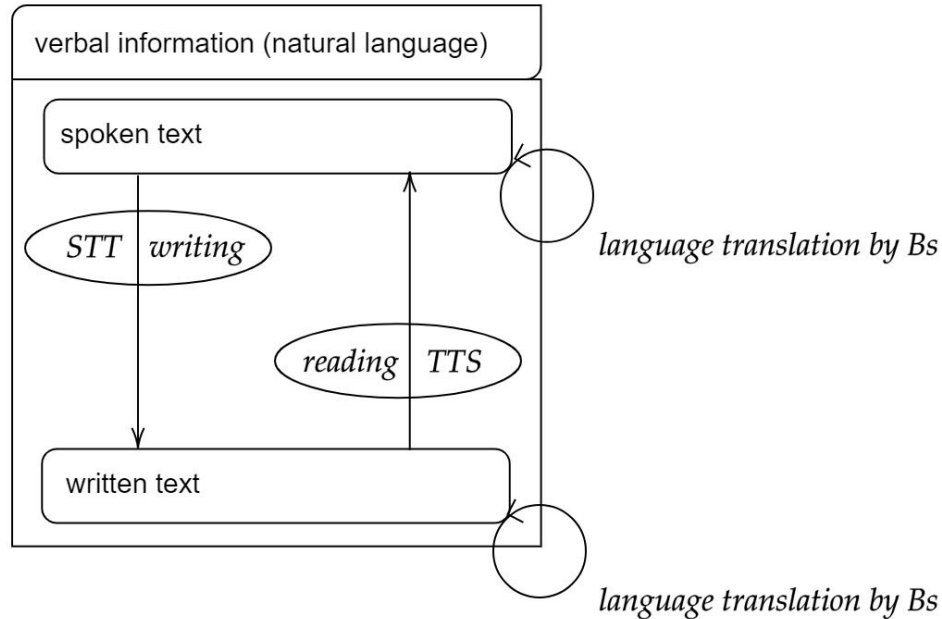
Why learn NLP?

- ▷ Many applications!
- ▷ Major contributor to growing AI market
- ▷ Demand is high
- ▷ Understanding how machines can learn language more efficiently and correctly
- ▷ Promote the diversity and preservation of Philippine languages



Introduction to language and natural language processing (NLP)

Building blocks of natural language



- characters
- morphemes
- words

To be discussed by the next set of slides on:
Words, morphology, and the lexicon

References

- ▷ Bird, S. 2008. *Defining a core body of knowledge for the introductory computational linguistics curriculum*. In Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics (TeachCL '08). Association for Computational Linguistics, USA, 27–35.
- ▷ Gleitman, L. 2005. *Language and thought*. Cambridge Handbook of Thinking and Reasoning.
- ▷ Jurafsky, D., and Martin, J. H. 2009. *Speech and Language Processing*, 2nd edition. Stanford University and University of Colorado at Boulder.
- ▷ Liu, S. 2020. AI market size 2018-2025. Retrieved from:
<https://www.statista.com/statistics/607716/worldwide-artificial-intelligence-market-revenues/>