

# Introduction to Natural Language Processing

Lecture-1

# Natural Language

Natural Language

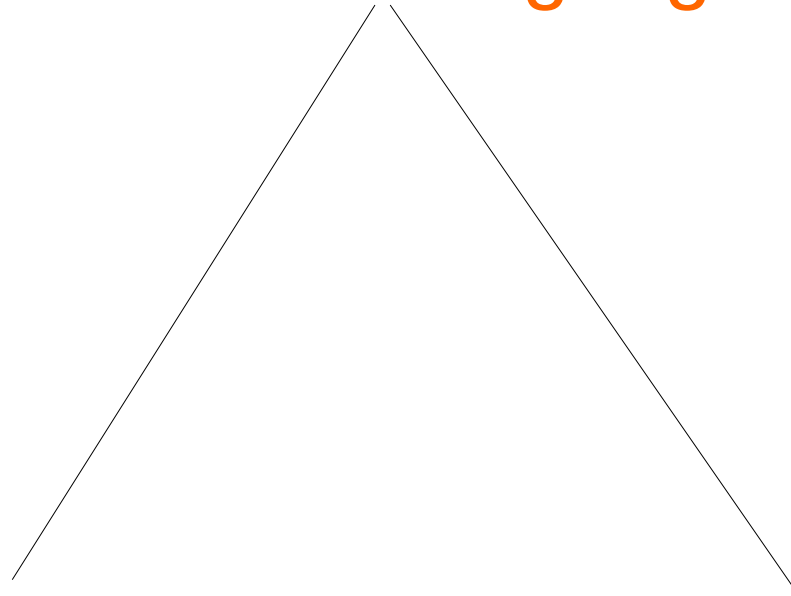
Text Mining

# Natural Language

Machine Learning/Deep  
Learning

Text Mining

Natural Language



Machine Learning/Deep  
Learning

Text Mining

# Introduction to Natural Language Processing

## Lecture-1

What is a Language?

What is a Natural Language?

What is Un-Natural Language?

# Artificial Language

```
try {
    cMessage = messageQueue.take();
    for (AsyncContext ac : queue) {
        try {
            PrintWriter acWriter = ac.getWriter();
            acWriter.println(cMessage);
            acWriter.flush();
        } catch (IOException e) {
            System.out.println("Error: " + e.getMessage());
        }
    }
} catch (InterruptedException e) {
    // ...
}
```

- append(CharSequence csq)
- append(char c)
- append(CharSequence csq, int start, int end)
- format(String format, Object... args)
- format(Locale l, String format, Object... args)
- printf(String format, Object... args)

(<https://netbeans.org/features/java/>)

```
def add5(x):
    return x+5

def dotwrite(ast):
    nodename = getNodeName()
    label=symbol.sym_name.get(int(ast[0]),ast[0])
    print '    %s [label="%s" % (nodename,label),
    if isinstance(ast[1], str):
        if ast[1].strip():
            print '= %s';' % ast[1]
        else:
            print ''
    else:
        print '];'
        children = []
        for n, child in enumerate(ast[1:]):
            children.append(dotwrite(child))
        print '    %s -> (' % nodename,
        for name in children:
            print '%s' % name,
```

(<http://noobite.com/learn-programming-start-with-python/>)



# Language

A **vocabulary** consists of a set of **words** ( $w_i$ )



A **text** is composed of a sequence of **words** from a **vocabulary**

A **language** is constructed of a set of all possible **texts**



(<http://www.old-engl.sh/language.php>)



## Beyond the genome

Studies of the epigenetic signatures of many healthy and diseased human tissues could provide crucial information to link genetic variation and disease.

**T**he Greek prefix *epi-* can signify upon, over, over, next, at, before, and after. Most of these could apply to its use in the term 'epigenetic' — particularly the last of these. It is some 18 years, almost to the day, that Nature published the draft sequence of the human genome. Since, in this issue, we publish results from a subsequent study on the same genome, modifications to the genome — epigenetic modifications — that crucially determine which genes are expressed by which cells, and when.

It is hard to think of any branch of human biology that has not been affected by the human genome sequence. In fact, the perhaps most important advance in our understanding of the past that genetics and genomics variation play in the normal functioning of a human body and in disease. But despite the progress, and despite the fact that the genome has been sequenced, many questions remain. Much remains to be understood about how genetic information is interpreted by the individual cells in our body.

This is where epigenetics comes in. Upon the genome, so the genome, over the genome — take your pick — epigenetics collectively describes changes in the regulation of gene expression that can be passed on to a cell progeny but are not due to changes in the nucleotide sequence of the gene.

Soon after the human genome sequence had been completed, it became clear that an epigenome — a map of the genome-wide modifications made to DNA and the proteins scaffold that supports it — would also be required. The task at hand was, as researchers like us, not trivial. Every cell in the body carries the same genome (with a few exceptions), but the epigenome changes with cell and tissue type.

Epigenetics is still an emerging science, but researchers are now building tools to study epigenetic changes in the genome in a systematic and genome-wide way. In 2012, Nature published the publication of the results of the ENCODE project, the aim of which was to identify all the functional elements encoded in the human genome by mapping epigenetic modifications (see [www.encodeproject.org](http://www.encodeproject.org)).

ENCODE was a pioneer in scale of effort and development of specialized analytical software, and has already had a tremendous impact on human genetics studies. But its direct application to the understanding of disease is still in its infancy. The project set out to generate and publicly share epigenomic data from many cells. From mature cells from a variety of different tissues from healthy people, and from patients with diseases such as cancer, and neurodegenerative and autoimmune diseases.

The main results of this vast project are published in this issue

starting on page 303, as well as in several other Nature Publishing Group journals.

Together, these fundamental aspects of epigenetics emerge how the epigenome affects gene expression, how the epigenome changes during normal development (that is, during normal development) and how it changes during disease.

The results emphasize the central role of epigenetic information in understanding these processes. Crucially, what emerges is that it is not just one or two types of modification that matter. Biology is rarely that simple. Instead, combinations of modifications produce gene activity in ways that a single type of modification does not. A causal link between epigenetic changes and disease has so far been hard to establish. Identifying such changes is necessary, however, if we are to understand the underlying disease mechanisms and design targeted treatments. With the new wealth of data, consistent alterations in the epigenome landscape could identify candidate genes and pathways for further follow-up. And tissue-specific studies of the epigenome of cell types relevant to a specific disease could indicate whether epigenetic changes have a role in disease progression, or only in its onset.

One reason that it has been difficult to relate some diseases to disruption in DNA function is that many of the key changes occur in poorly understood regions of the genome, usually outside those parts that code for proteins. Epigenetic maps such as those published today should help scientists to navigate this poorly charted landscape. By overlaying these maps, maps to relevant cell types, researchers can determine, for example, whether an epigenetic change associated with a given disease is in a region of the genome that regulates gene activity. If so, then this may provide a possible lead to be explored.

Cancer is often called the disease of the genome, but the genome does not exist, or operate, in splendid isolation. Of all diseases, cancer has been linked most consistently to epigenetic alterations. Scientists have long suspected that epigenetic misregulation affects the genomic location of the mutations that provide cancer. The new findings suggest that this is true, and they go further. They show that the epigenome of a cancer cell carries a fingerprint of the cell type that originated the cancer. This is a crucial information, especially for cancers in complex tissues such as the brain, which cannot presently be traced to their original cell type.

In human diseases, the genome and epigenome operate together. Tackling disease using information on the genome alone has been like trying to work with one hand tied behind the back. The new views of epigenetic data from the other hand. It will now provide all the answers, that it could help researchers decide which questions to ask.

19 FEBRUARY 2013 | VOL 350 | NATURE | 315  
© 2013 Macmillan Publishers Limited. All rights reserved.

([http://www.nature.com/polopoly\\_fs/1.16929!/menu/main/topColumns/topLeftColumn/pdf/518273a.pdf](http://www.nature.com/polopoly_fs/1.16929!/menu/main/topColumns/topLeftColumn/pdf/518273a.pdf))

<http://learnenglish.britishcouncil.org/en/vocabulary-games>)

# A Tool to Shape Minds

- Edward Sapir & Benjamin Lee Whorf: language determines how one thinks
- Lev Vygotsky: language guides the child's cognitive growth
- Katherine Nelson: language acts as the medium through which the mind becomes part of a culture



# A Tool to Shape Minds

- We not only speak, but also listen.
- The listening is no less important than the speaking: the speaking expresses our mind, but the listening shapes our mind.
- Language creates minds.

# Redundant and Inefficient

- On average western languages are about 50% redundant: we would not lose any expressive power if we gave up 50% of our dictionary. We can guess the meaning of most sentences from a fragment of them.
- Human communication is wildly inefficient: two computers can simply exchange in a split second an image without any loss of information, whereas a human must describe to another human the image in a lengthy way and will certainly miss some details



# Rationalist and Empiricist Approaches to Language

## Empiricist Approaches

- Belief that a significant part of the knowledge in the human mind is not derived by the senses but is fixed in advance, presumably by genetic inheritance. (Chomsky)
  - Argues innate structure because of poverty of the stimulus
  - it is difficult to see how children can learn something as complex as a natural language from the limited input (of variable quality and interpretability) that they hear during their early years

•

# Rationalist Approach

- Key parts of language are innate - hardwired in the brain at birth as part of the human genetic inheritance

# Grammar



- Noam Chomsky:
  - The number of sentences in a language is potentially infinite, but there is a finite system of rules that defines which sentences can potentially be built
  - You have never read a sentence with these exact words before but (hopefully!) you understand the meaning of what I just wrote

# Grammar

- Noam Chomsky:
  - The logical formalism used to prove mathematical theorems can be employed to express the grammar of a language
  - The grammar of a language “is” the specification for the entire language

$S \rightarrow NP VP$   
 $VP \rightarrow VP PP$   
 $VP \rightarrow V NP$   
 $VP \rightarrow eats$   
 $PP \rightarrow P NP$   
 $NP \rightarrow Det N$   
 $NP \rightarrow she$   
 $V \rightarrow eats$   
 $P \rightarrow with$   
 $N \rightarrow fish$   
 $N \rightarrow fork$   
 $Det \rightarrow a$



# Grammar

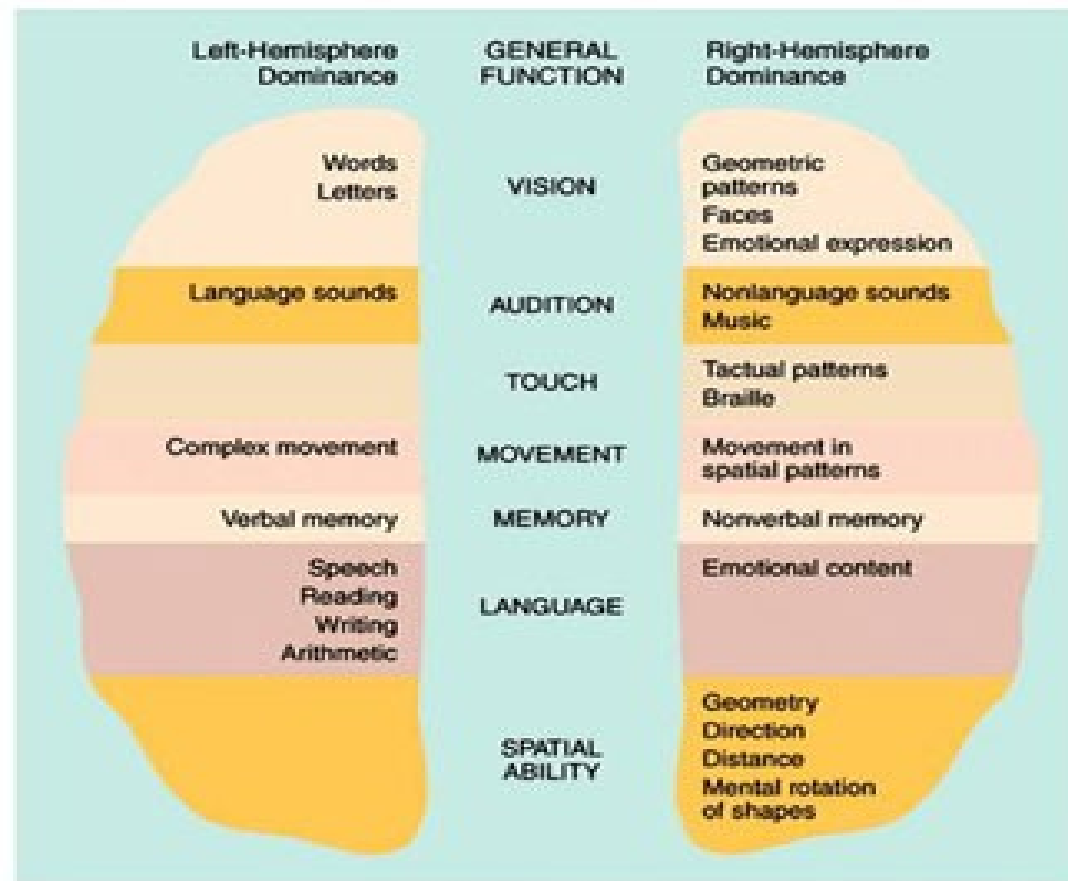
- Chomsky:
  - Children do not learn, as they do not make any effort. Language "happens" to a child.
  - We are born with some innate knowledge of what a grammar is and how it works (a "universal grammar")
  - Then experience determines which specific language (i.e., grammar) we will learn.
  - We are predisposed to learn a language the same way we are predisposed to learn to eat
  - Language acquisition is not only possible: it is virtually inevitable

# Grammar

- “Xgewut is not a meaningful word” is a correct English sentence. What makes a sentence correct even when it contains a word that does not exist?

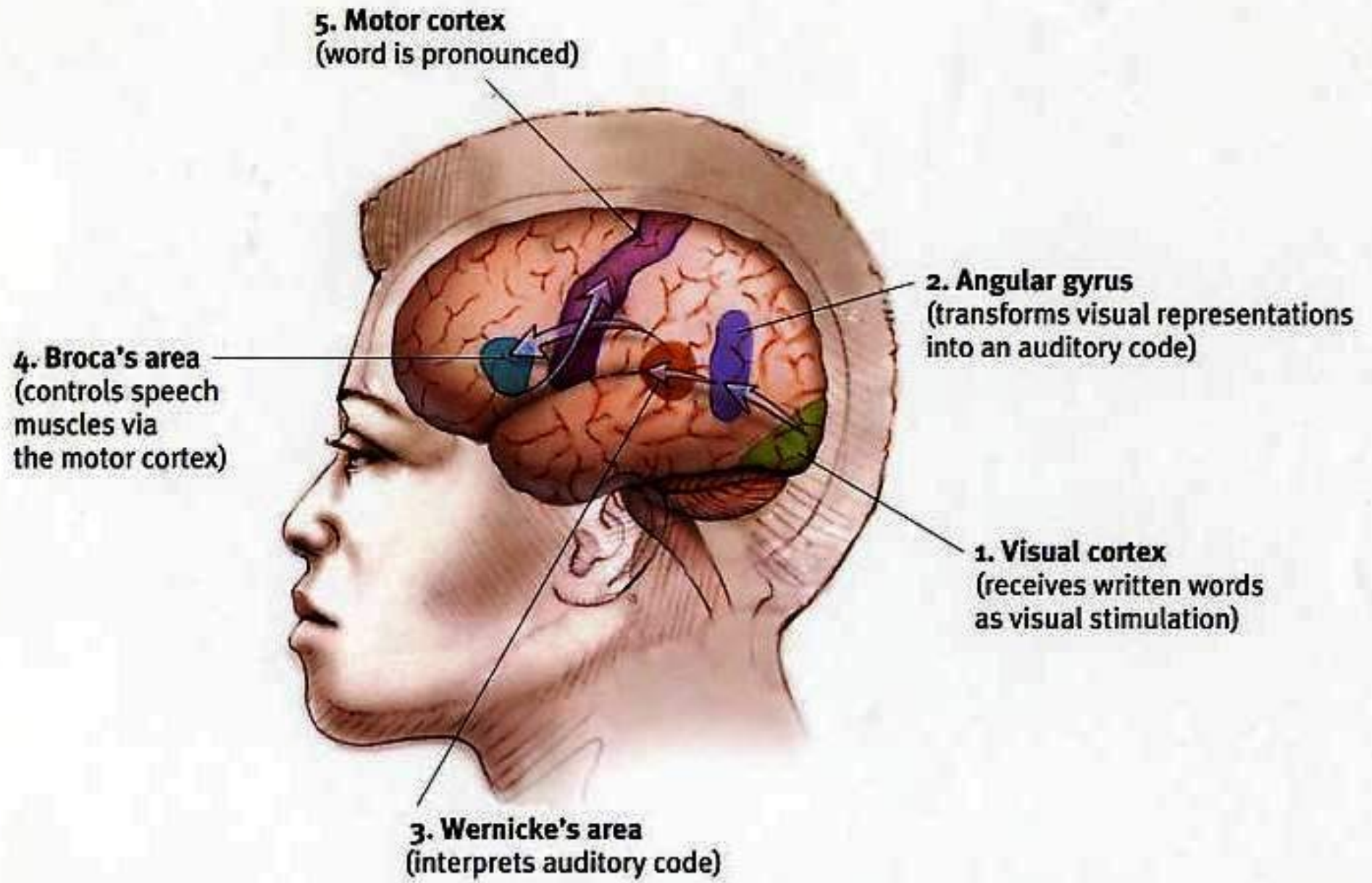
# Abilities that display cerebral lateralization of function

## ► Abilities That Display Cerebral Lateralization of Function



# • Biological Foundations

- Language is predominantly associated with the **left hemisphere** of the brain.
  - Wernicke's Area
    - Affects comprehension in speech that is heard and text that is read.
  - Broca's Area
    - Affects the production of language through speaking or writing.
- Individual differences in language ability are due to **genetics**.
- Critical Periods for Language Development
  - Lenneberg proposed that language must be acquired **before adolescence**.
- Speed of Acquisition Relative to the Amount of Input for Language Development
  - **Children acquire language with little intervention**



# Ambiguity in Language

*"Shi shi shi shi shi shi shi shi shi shi shi  
shi shi"*

("the master is fond of licking lion spittle")

(Chinese tongue-twister)

# Ambiguity in Language

Buffalo buffalo Buffalo buffalo buffalo buffalo Buffalo buffalo.

- a. a city named Buffalo. This is used as a noun adjunct in the sentence;
- n. the noun buffalo (American bison), an animal, in the plural (equivalent to "buffaloes" or "buffalos"), in order to avoid articles.
- v. the verb "buffalo" meaning to outwit, confuse, deceive, intimidate, or baffle.

[Those] buffalo(es) from Buffalo [that are intimidated by] buffalo(es) from Buffalo intimidate buffalo(es) from Buffalo.

Bison from Buffalo, New York, who are intimidated by other bison in their community, also happen to intimidate other bison in their community.

The buffalo from Buffalo who are buffaloed by buffalo from Buffalo, buffalo (verb) other buffalo from Buffalo. Buffalo buffalo (main clause subject) [that] Buffalo buffalo (subordinate clause subject) buffalo (subordinate clause verb) buffalo (main clause verb) Buffalo buffalo (main clause direct object).

[Buffalo from Buffalo] that [buffalo from Buffalo] buffalo, also buffalo [buffalo from Buffalo].

- **Will, will Will will Will Will's will?** –
  - Will (a person), will (future tense helping verb) Will (a second person) will (bequeath) [to] Will (a third person) Will's (the second person) will (a document)? (Someone asked Will 1 directly if Will 2 plans to bequeath his own will, the document, to Will 3.)
- 
- **Police police Police police police police Police police**
  - Cops from Police, Poland, whom cops from Poland patrol, patrol cops from Poland.
- 
- **Rose rose to put rose roes on her rows of roses.**
  - Rose [a person] rose [stood] to put rose [pink-colored] roes [fish eggs as fertilizer] on her rows of roses [flower].



- James while John had had had had had had had had had had had had a better effect on the teacher[3] –
  - With punctuation: "James, while John had had 'had', had had 'had had'. 'Had had' had had a better effect on the teacher", or James, while John had had 'had had', had had 'had'. 'Had had' had had a better effect on the teacher
- That that is is that that is not is not is that it it is – Grammatically corrected as: "That that is, is. That that is not, is not. Is that it? It is".

- Can can can can can can can can can. – "Examples of the can-can dance that other examples of the same dance are able to outshine, or figuratively to put into the trashcan, are themselves able to outshine examples of the same dance". It could alternatively be interpreted as a question, "Is it possible for examples of the dance that have been outshone to outshine others?" or several other ways.
- 
- Martin Gardner offered the example: "Wouldn't the sentence 'I want to put a hyphen between the words Fish and And and And and Chips in my Fish-And-Chips sign' have been clearer if quotation marks had been placed before Fish, and between Fish and and, and and and And, and And and and, and and and And, and And and and, and and and Chips, as well as after Chips?"

- Syntactic ambiguity

- We saw her duck.
- 
- One morning I shot an elephant in my pajamas.  
How he got in my pajamas, I don't know.
- 
- Time flies like an arrow; fruit flies like a banana

# Syntactic ambiguity, incrementality, and local coherence

- The horse raced past the barn fell.
- The coach smiled at the player tossed the frisbee (by the opposing team).[7]
- While the man was hunting the deer ran through the forest.

# • Embedding

- The rat the cat the dog bit chased escaped.
- The editor authors the newspaper hired liked laughed.
- The man who the boy who the students recognized pointed out is a friend of mine.

# Anaphora

- "He went to bed" (who?)
- "Today I wrote this sentence" (which day?)
- "Here it is cold" (where?)

# Metaphor

- “Her marriage is a nightmare”
- “My room is a jungle”
- “He is a snake”
- “This job is a piece of cake”
- “Time is money”

# Pragmatics

- What are the speaker's motif and goal?
- Semantics can account for the meaning of the sentence "do you know what time it is?", but not for the fact that an answer is required (the speaker's intention is to learn what time it is)



# Jokes

- What is a joke?
- Why do we tell jokes?
- What is in a joke?
- In order to understand a joke one must master the whole power of the language

# Evolution of Language

- Formal writing
- Informal writing
- Texting
- Code-mixing

What is Natural Language Processing?

# Natural Language Processing (NLP)

- The ability of a computer program to understand human language as it is spoken.

# Why NLP is Hard?

# • Aspects of language processing

- Word, lexicon: lexical analysis
  - Morphology, word segmentation
- Syntax
  - Sentence structure, phrase, grammar, ...
- Semantics
  - Meaning
  - Execute commands
- Discourse analysis
  - Meaning of a text
  - Relationship between sentences (e.g. anaphora)
-

# Tools

- WordNet
- CoreNLP from Stanford group.
- NLTK, the most widely-mentioned NLP library for Python.
- TextBlob, a user-friendly and intuitive NLTK interface.
- Gensim, a library for document similarity analysis.
- SpaCy, an industrial-strength NLP library built for performance.