Introduction to Natural Language Processing

Lecture-1

Natural Language

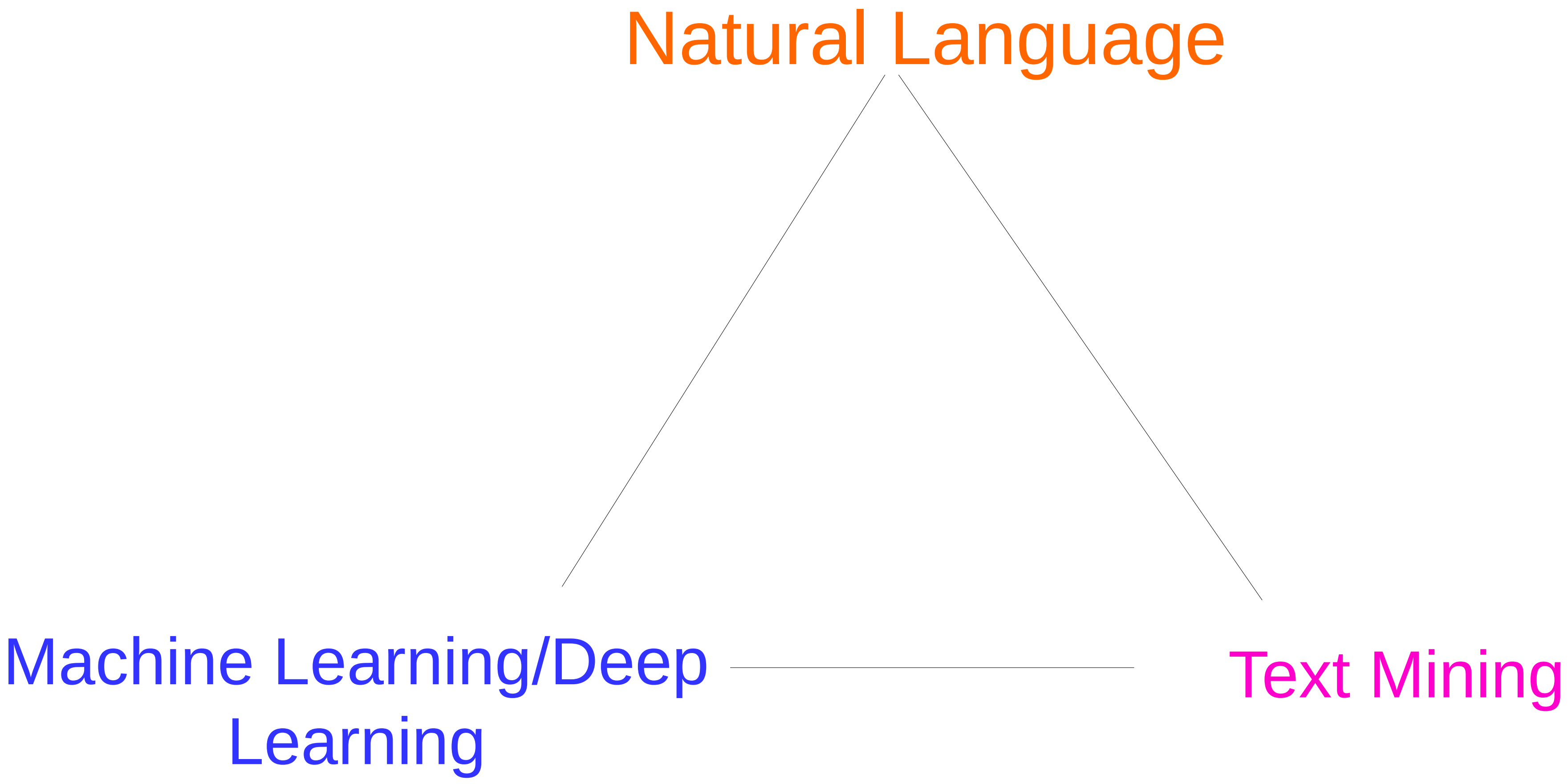
Natural Language

Text Mining

Natural Language

Machine Learning/Deep Learning

Text Mining



Natural Language

Machine Learning/Deep Learning

Text Mining

Introduction to Natural Language Processing

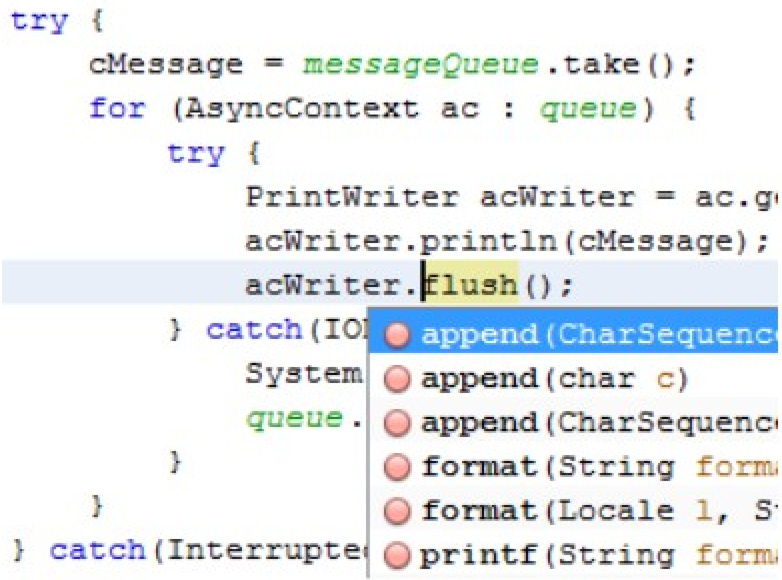
Lecture-1

What is a Language?

What is a Natural Language?

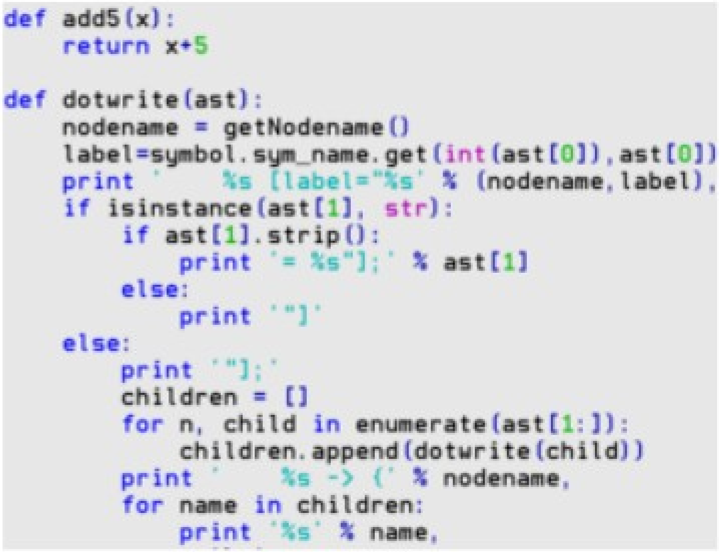
What is Un-Natural Language?

Artificial Language



```
try {
    cMessage = messageQueue.take();
    for (AsyncContext ac : queue) {
        try {
            PrintWriter acWriter = ac.g
            acWriter.println(cMessage);
            acWriter.flush();
        } catch(IO
            System
            queue.
        }
    }
} catch(Interrupte
```

(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,
```

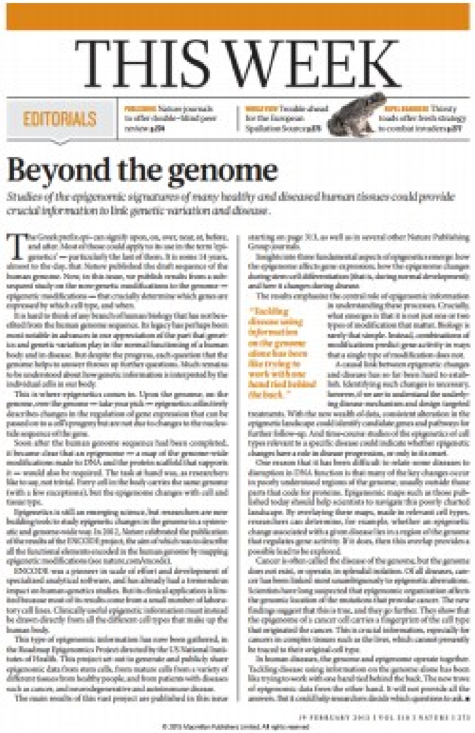
(http://noobite.com/learn-programming-start-with-python/)

Language

A vocabulary consists of a set of words (w_i)

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

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



THIS WEEK

EDITORIALS

Beyond the genome

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

[illegible]

"Tackling disease using information on the genome alone has been like trying to work with one hand tied behind the back."

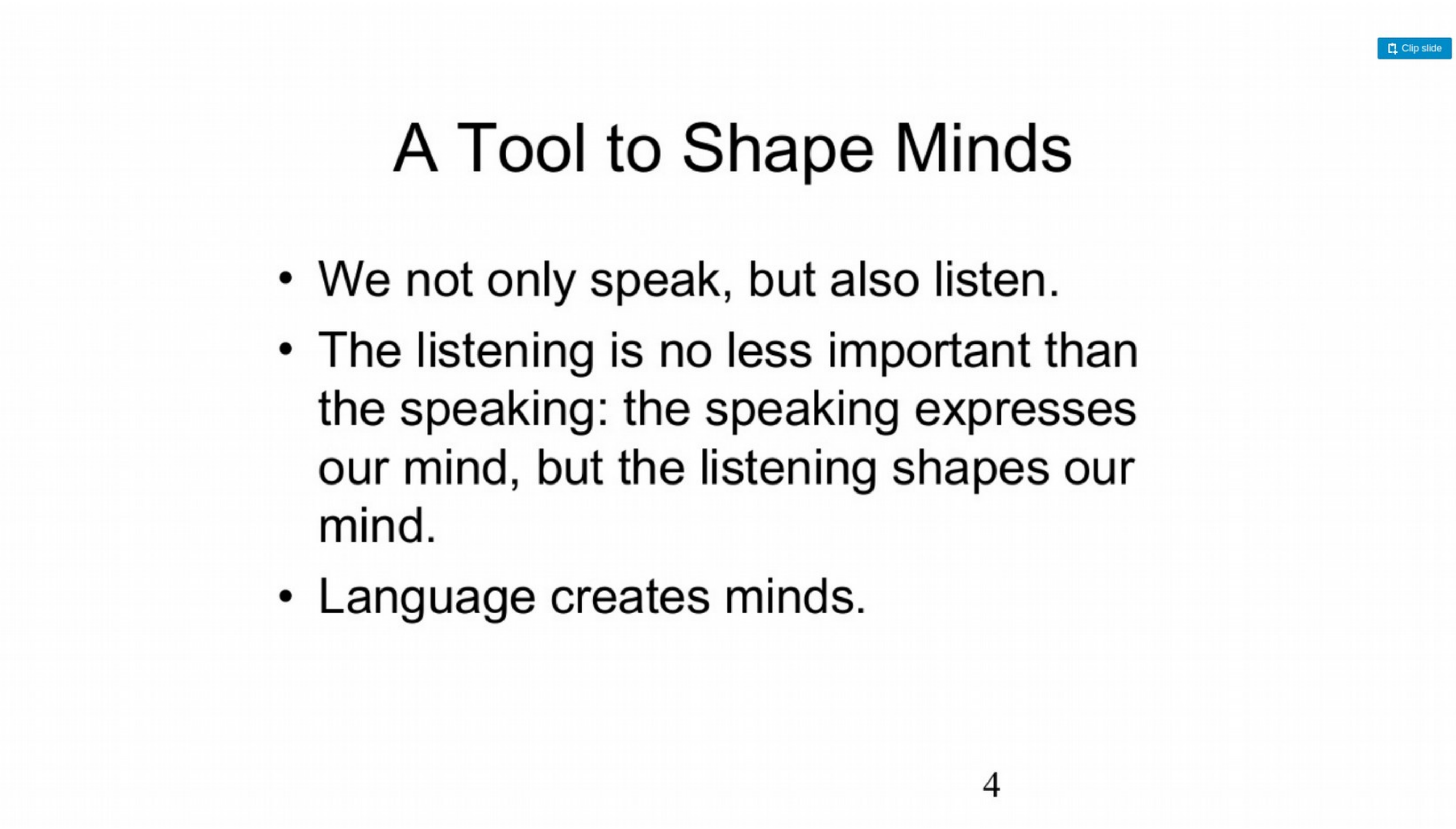
(http://www.nature.com/polopoly_fs/1.16929!/menu/main/topColumns/topLeftColumn/pdf/518273a.pdf)

A language is constructed of a set of all possible texts

(http://www.old-engli.sh/language.php)

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

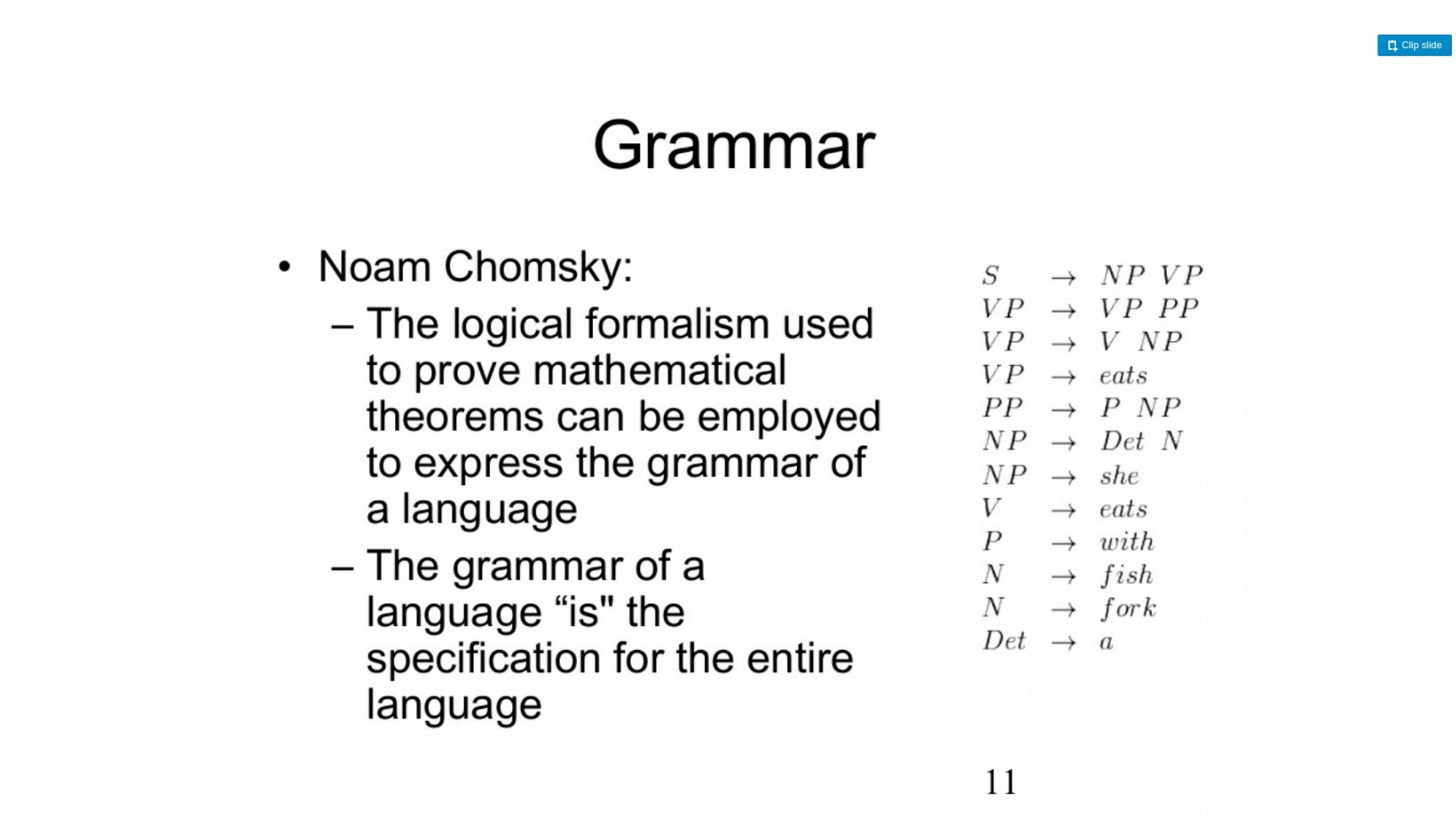
- Belief that a significant part of the knowledge in the human mind is not derived by the senses but is fixed in advance, presumably by genetic inheritance. (Chomsky)
    - Argues innate structure because of poverty of the stimulus
    - it is difficult to see how children can learn something as complex 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

$$
\begin{array}{lcl}
S & \rightarrow & NP\ VP \\
VP & \rightarrow & VP\ PP \\
VP & \rightarrow & V\ NP \\
VP & \rightarrow & eats \\
PP & \rightarrow & P\ NP \\
NP & \rightarrow & Det\ N \\
NP & \rightarrow & she \\
V & \rightarrow & eats \\
P & \rightarrow & with \\
N & \rightarrow & fish \\
N & \rightarrow & fork \\
Det & \rightarrow & a
\end{array}
$$

Grammar

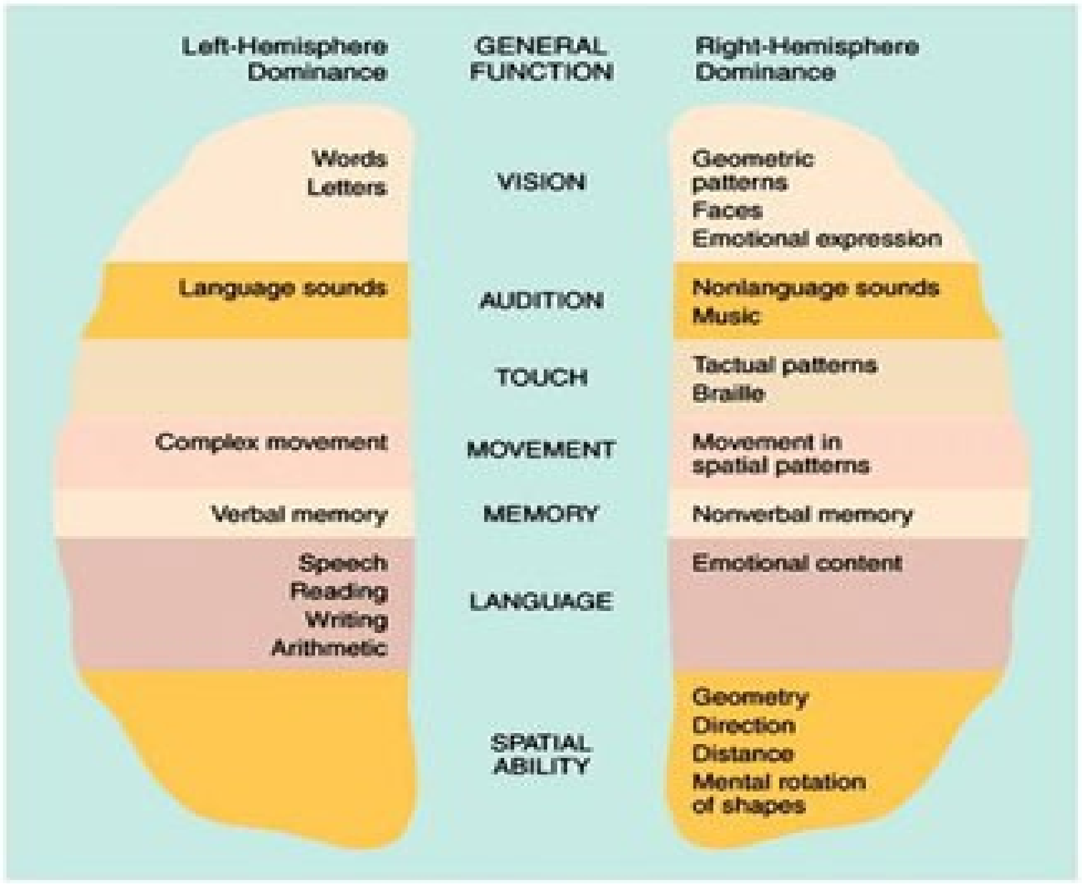
- Chomsky:
  - Children do not learn, as they do not make any effort. Language "happens" to a child.
  - We are born with some innate knowledge of what a grammar is and how it works (a “universal grammar”)
  - Then experience determines which specific language (i.e., grammar) we will learn.
  - We are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

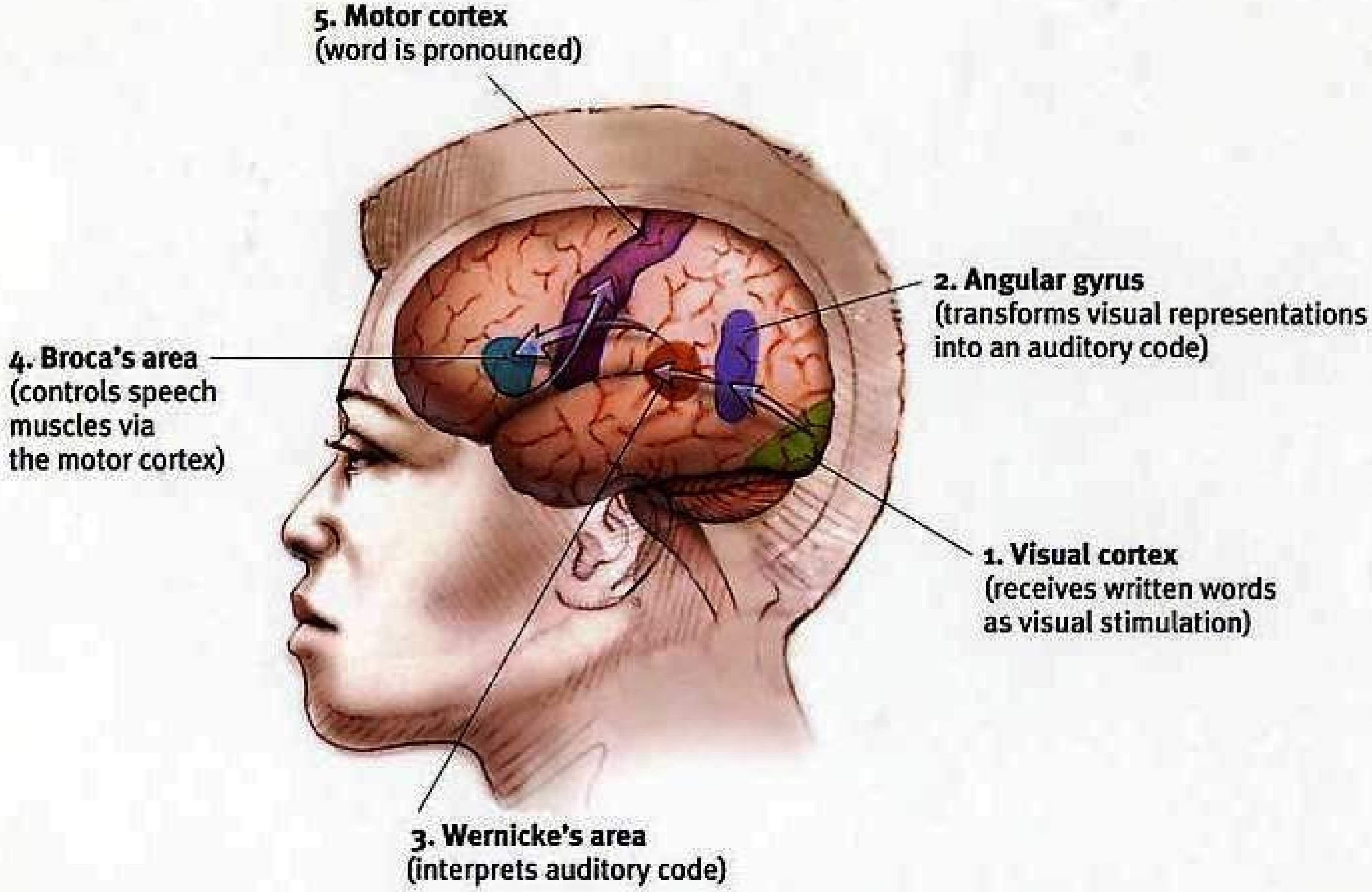
► **Abilities That Display Cerebral Lateralization of Function**



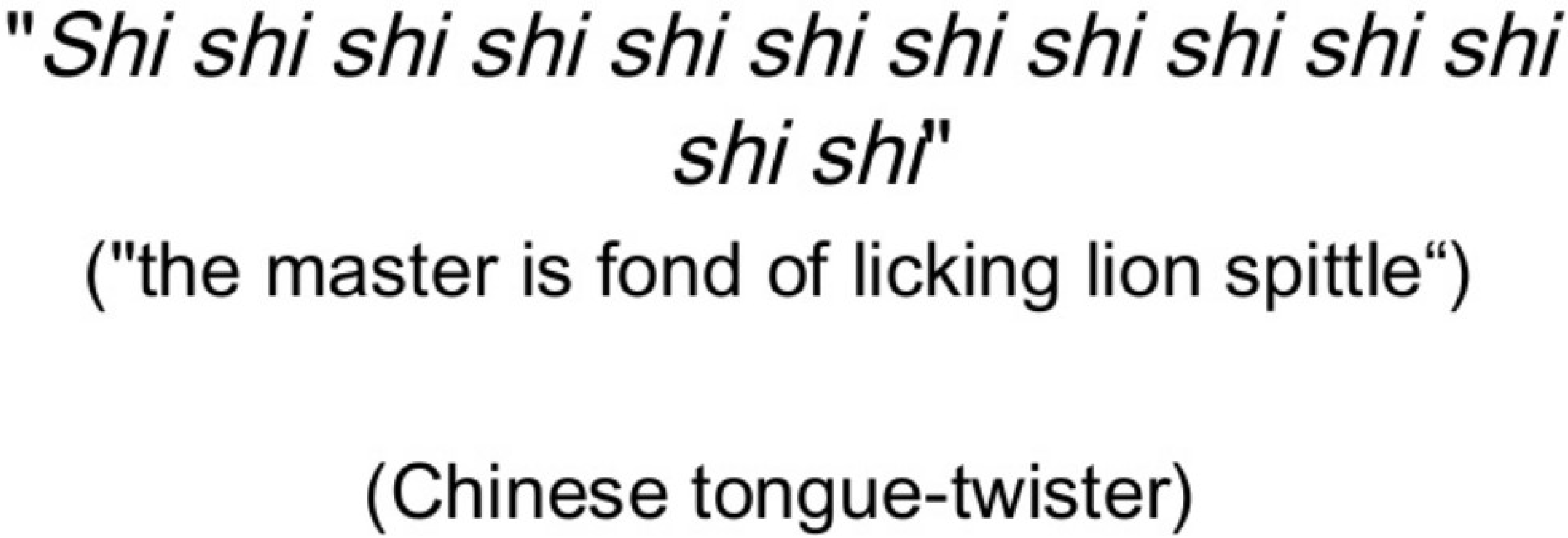
Left-Hemisphere Dominance	GENERAL FUNCTION	Right-Hemisphere Dominance
Words Letters	VISION	Geometric patterns Faces Emotional expression
Language sounds	AUDITION	Nonlanguage sounds Music
	TOUCH	Tactual patterns Braille
Complex movement	MOVEMENT	Movement in spatial patterns
Verbal memory	MEMORY	Nonverbal memory
Speech Reading Writing Arithmetic	LANGUAGE	Emotional content
	SPATIAL ABILITY	Geometry Direction Distance Mental rotation of shapes

• 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


5. Motor cortex
(word is pronounced)
2. Angular gyrus
(transforms visual representations
into an auditory code)
4. Broca's area
(controls speech
muscles via
the motor cortex)
1. Visual cortex
(receives written words
as visual stimulation)
3. Wernicke's area
(interprets auditory code)

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 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 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?
- I 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.