LECTURER: Nghia Duong-Trung

DATA UTILIZATION

TOPIC OUTLINE

TOPIC OUTLINE

Problem-Solving	5.2
Decision Support	6.1
Data Security and Data Protection	6.2

UNIT 3

NATURAL LANGUAGE PROCESSING (NLP)

STUDY GOALS

On completion of this unit, you will have learned ...

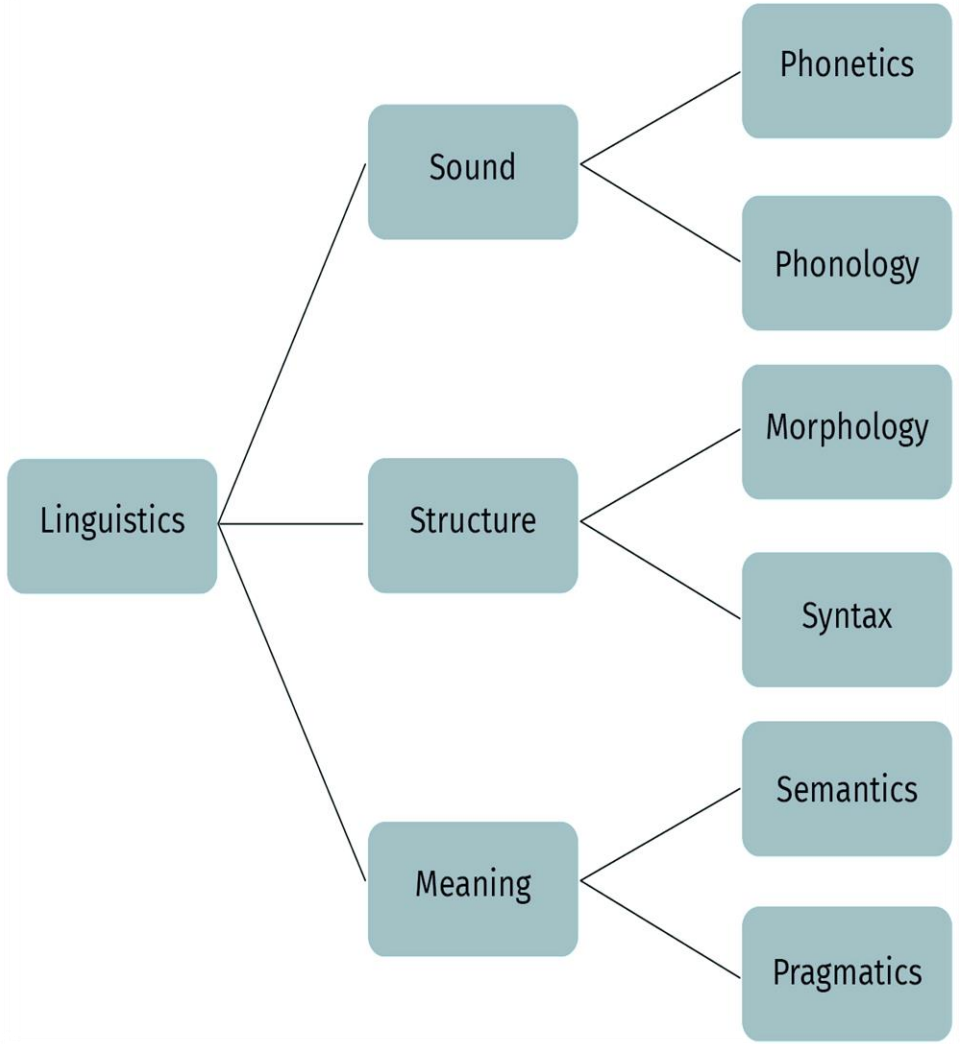
- … concepts of natural languages.
- … the three generations of natural language processing (NLP).
- … underlying concepts and models of speech recognition.
- … applications of NLP in social media analysis.
- … concepts of and techniques for different types of analyses, such as lexical, syntactical,and semantic analysis.

EXPLAIN SIMPLY

1. How can NLP analyze and understand a text and summarize it?
2. What are the levels of speech recognition?
3. How can NLP analyze social media data and what information is extracted from them?

LINGUISTICS

- Linguistics is the study of the nature of language and communication including phonetics, grammar, and words.
- Language is a dynamic and living phenomenon that evolved and changed over time.
- NLP is the procedure of processing languages or linguistics.



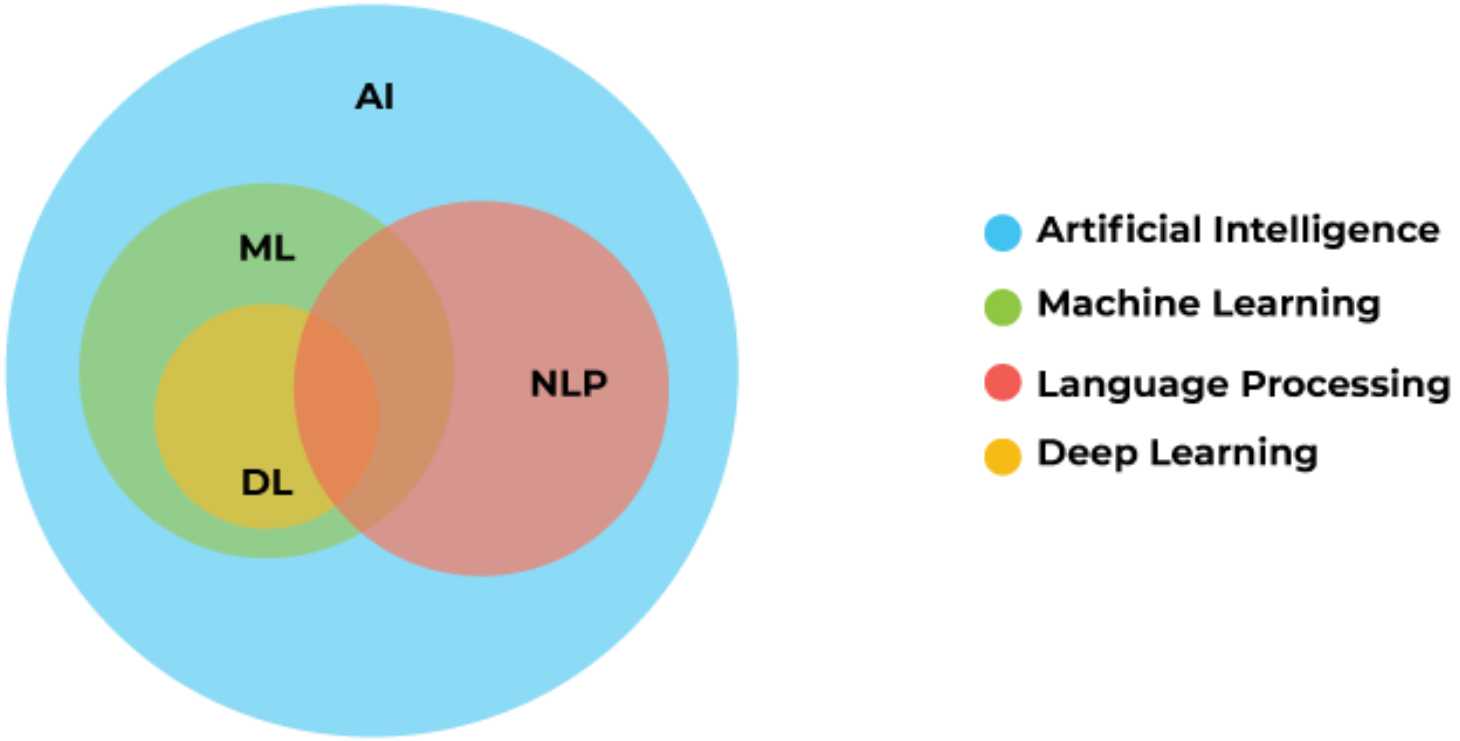
Source of the graphic: Course Book DLMBBD01, p. 65

- NLP has three main areas
  - natural language generation
  - natural language understanding
  - speech recognition

- NLP techniques can be categorized into three generations:
  - classical generation
  - statistical and empirical generation
  - deep learning

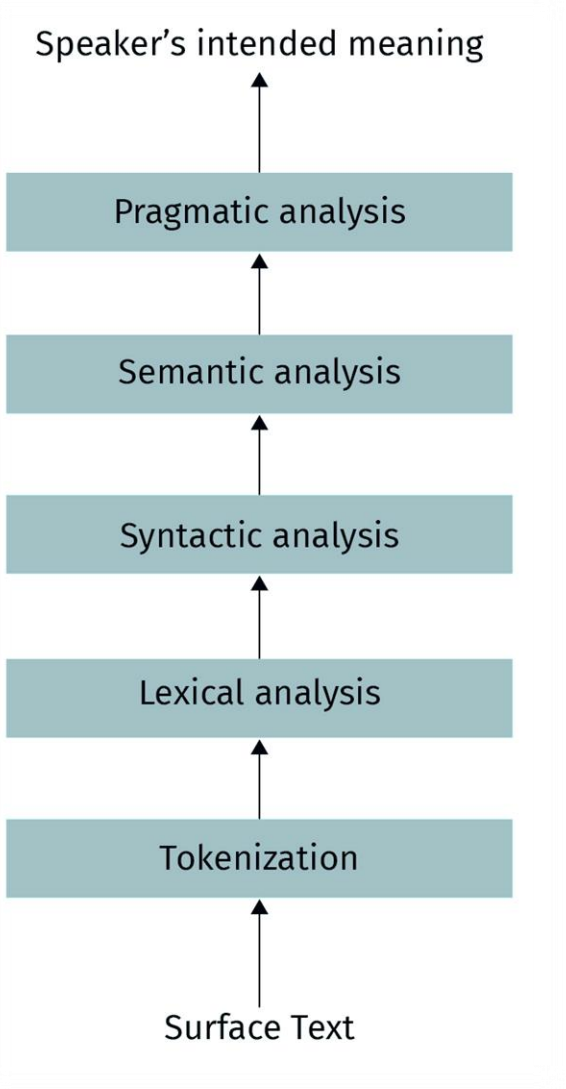
NLP IN ML AND AI

Natural Language Processing is a form of AI that gives machines the ability to not just read, but to understand and interpret human language. With NLP, machines can make sense of written or spoken text and perform tasks including speech recognition, sentiment analysis, and automatic text summarization.

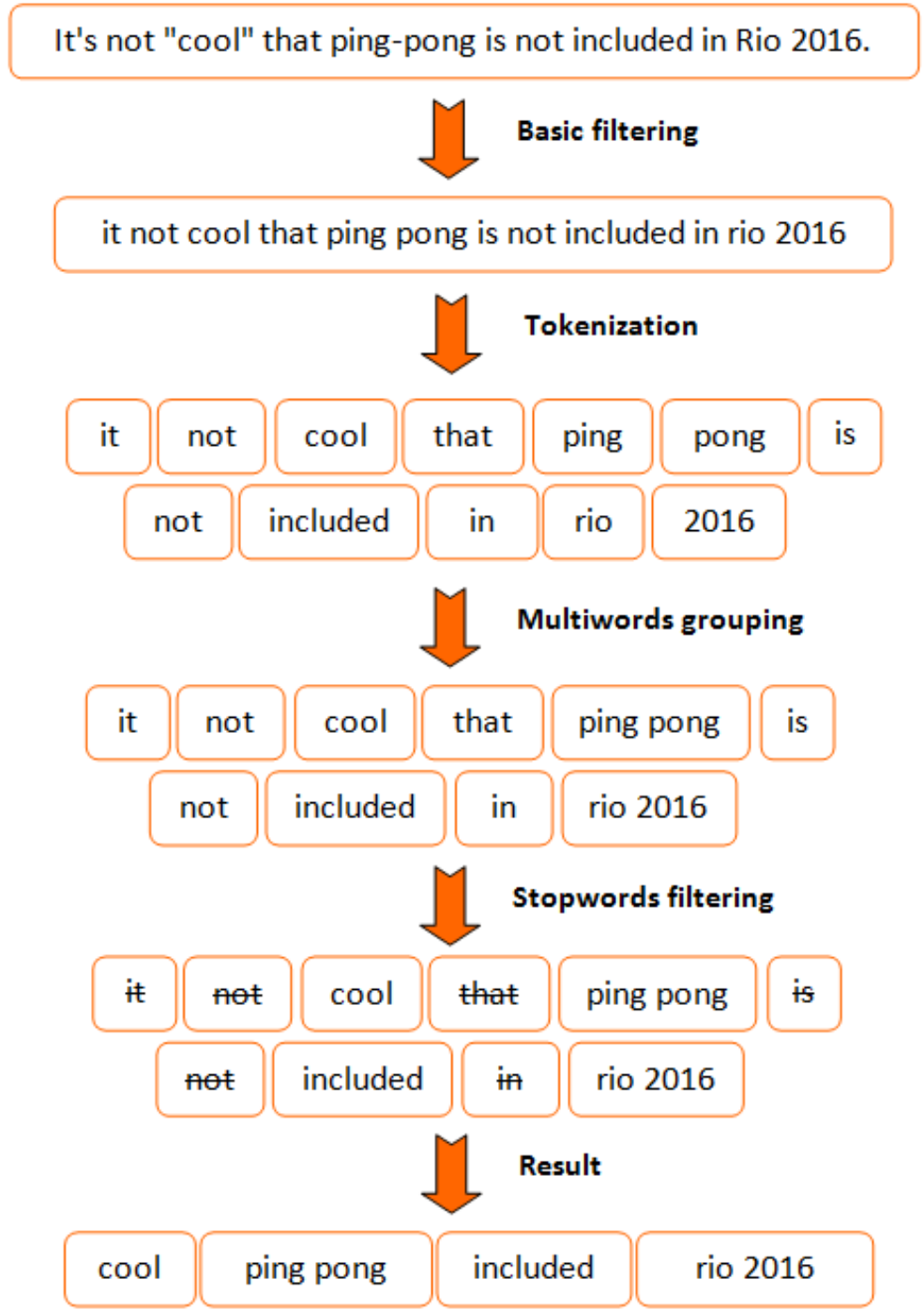


- The first step of NLP: breaking up a text into sentences and words
  - tokenization
  - sentence segmentation

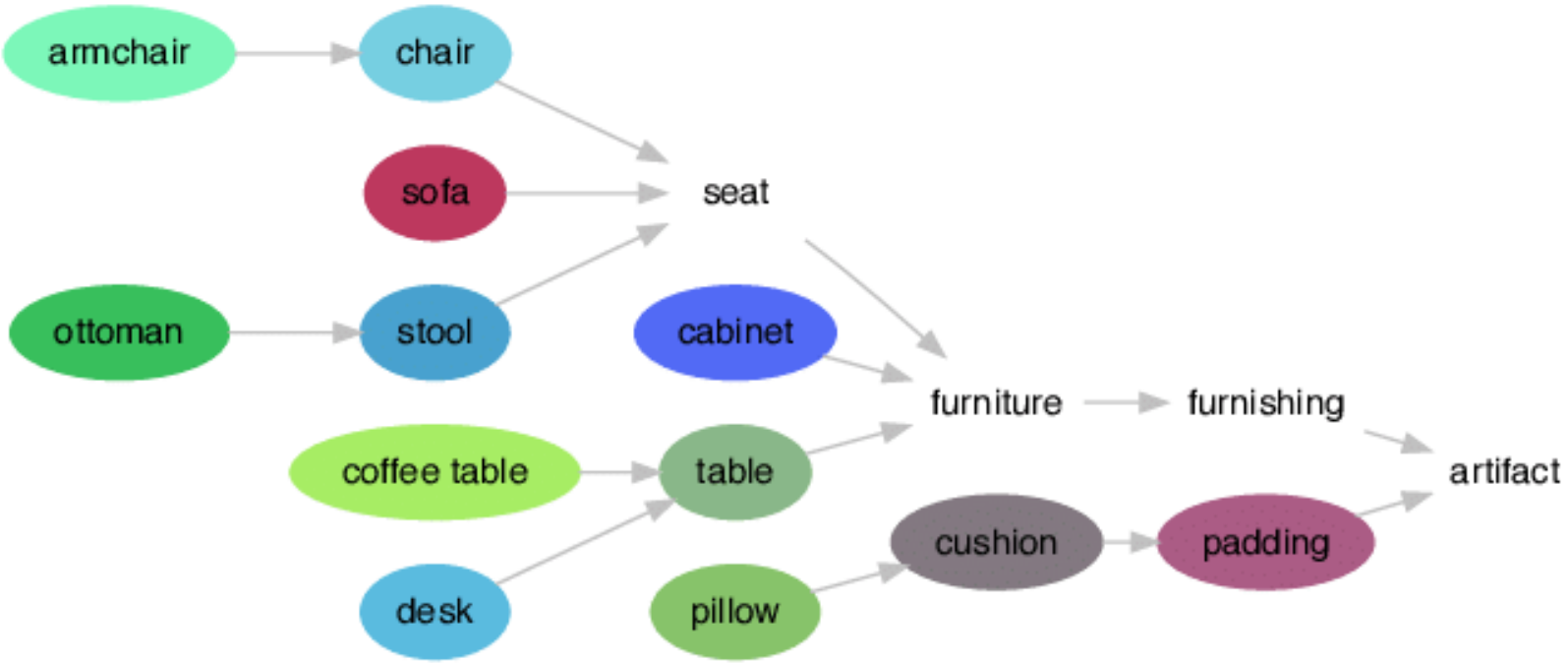
Attention: Some languages like Chinese do not share an easy delimited tokenization.



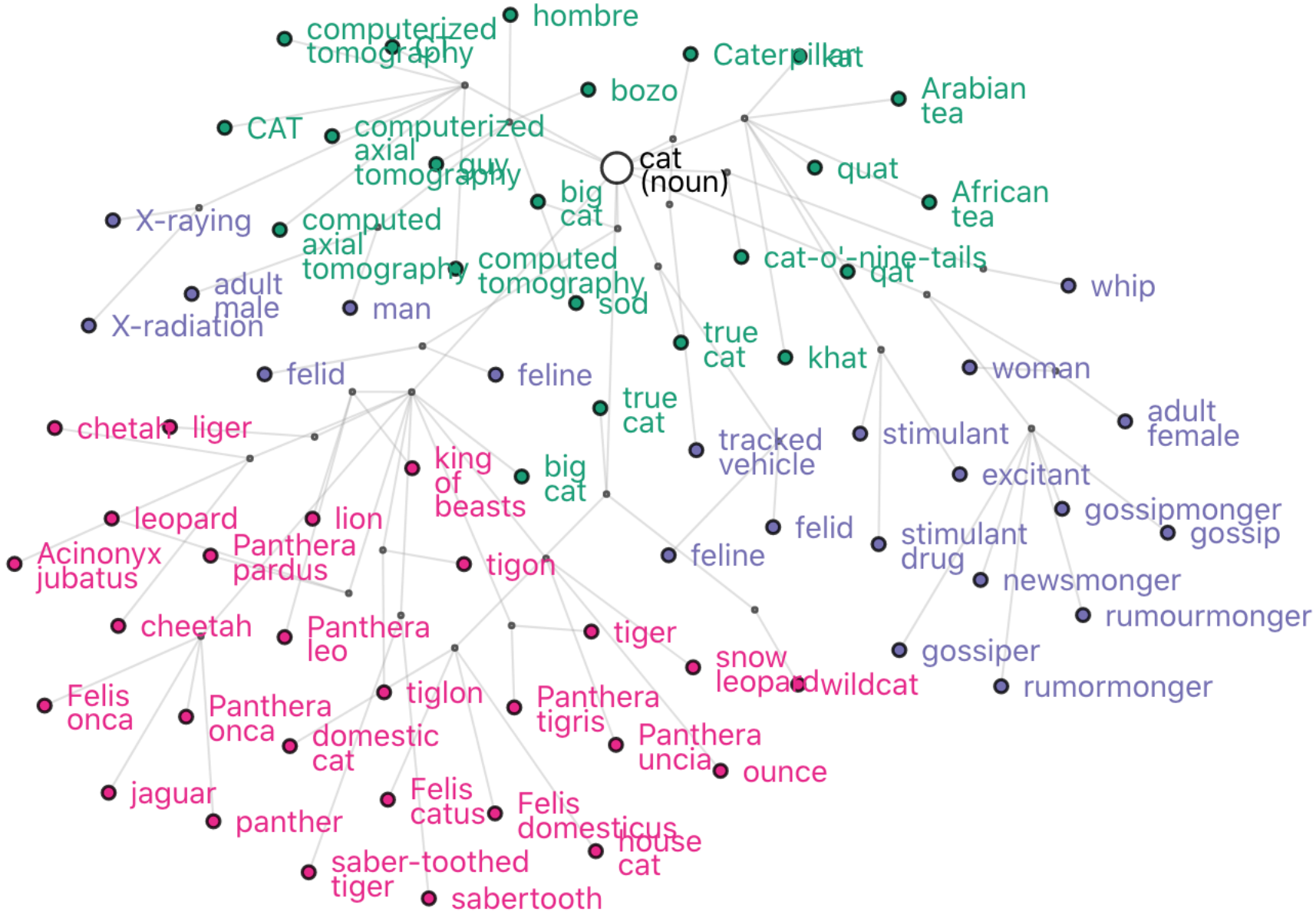
Source of the graphic: Indurkhya & Damerau, 2010.


It's not "cool" that ping-pong is not included in Rio 2016.
Basic filtering
it not cool that ping pong is not included in rio 2016
Tokenization
it not cool that ping pong is not included in rio 2016
Multiwords grouping
it not cool that ping pong is not included in rio 2016
Stopwords filtering
it not cool that ping pong is not included in rio 2016
Result
cool ping pong included rio 2016

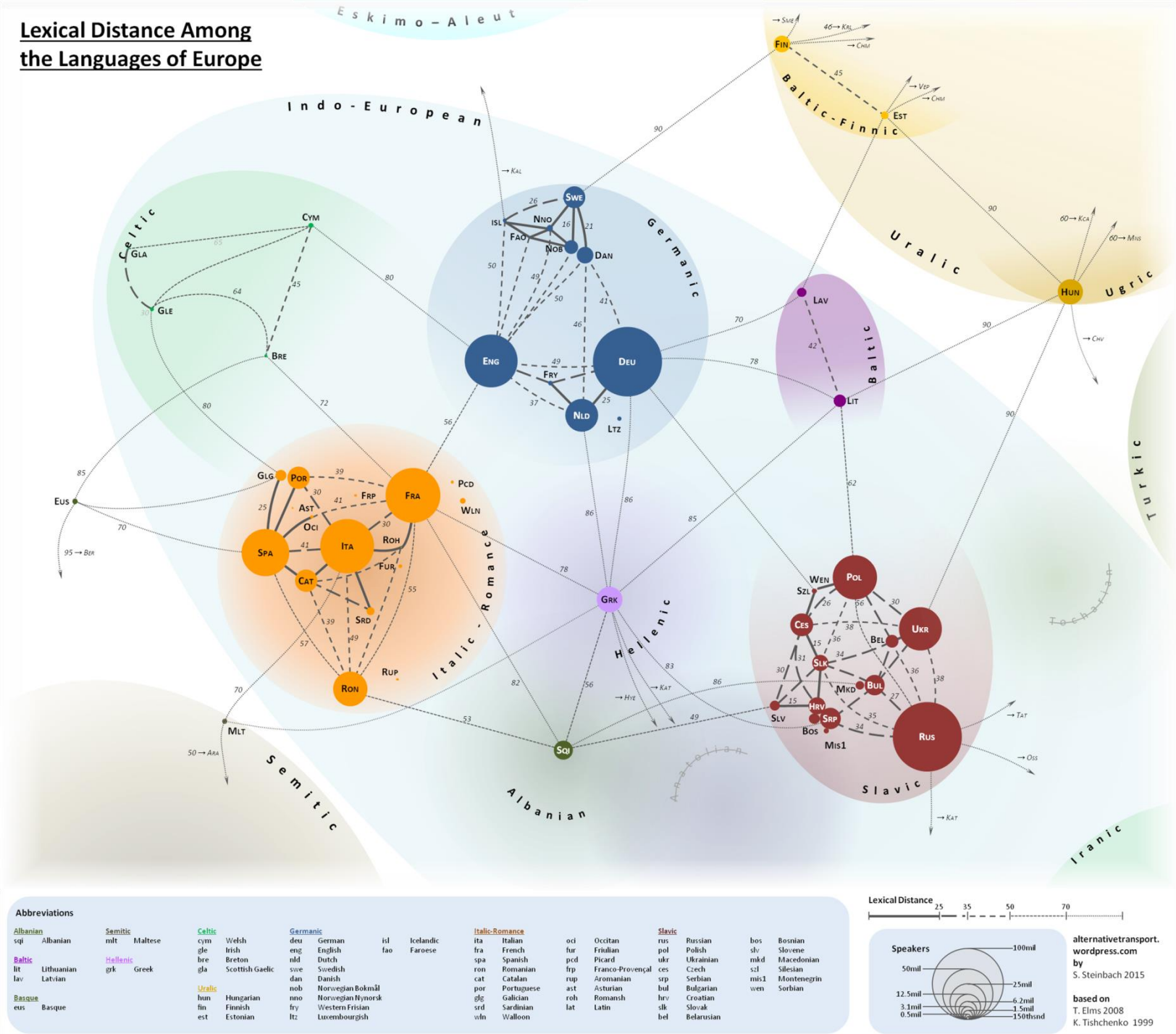
LEXICAL ANALYSIS



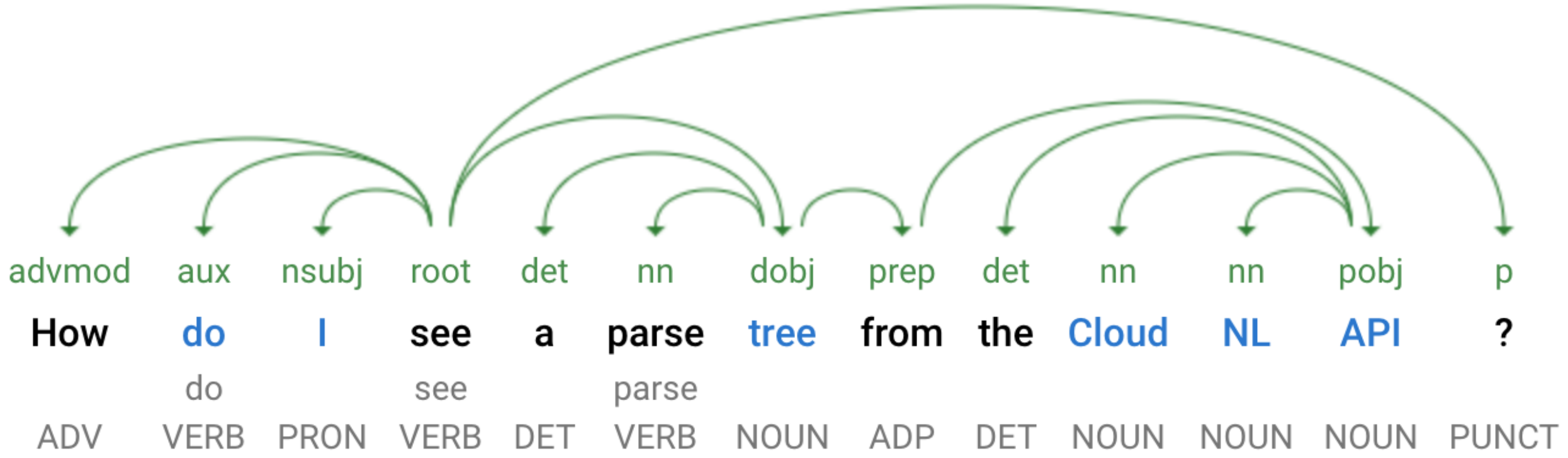
LEXICAL ANALYSIS



LEXICAL ANALYSIS

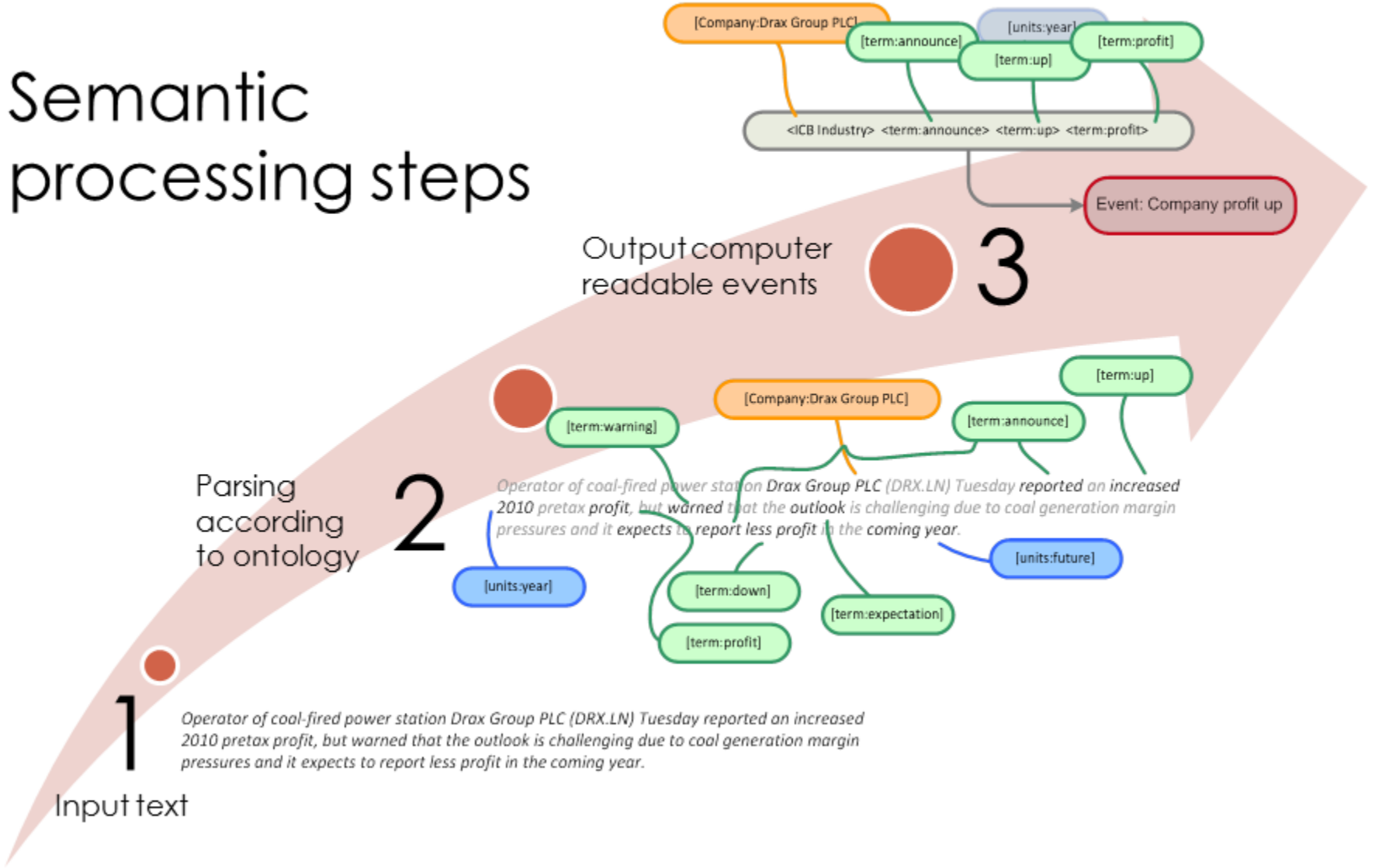


SYNTACTIC ANALYSIS



ViewerPro
semantic news analysis

Semantic processing steps



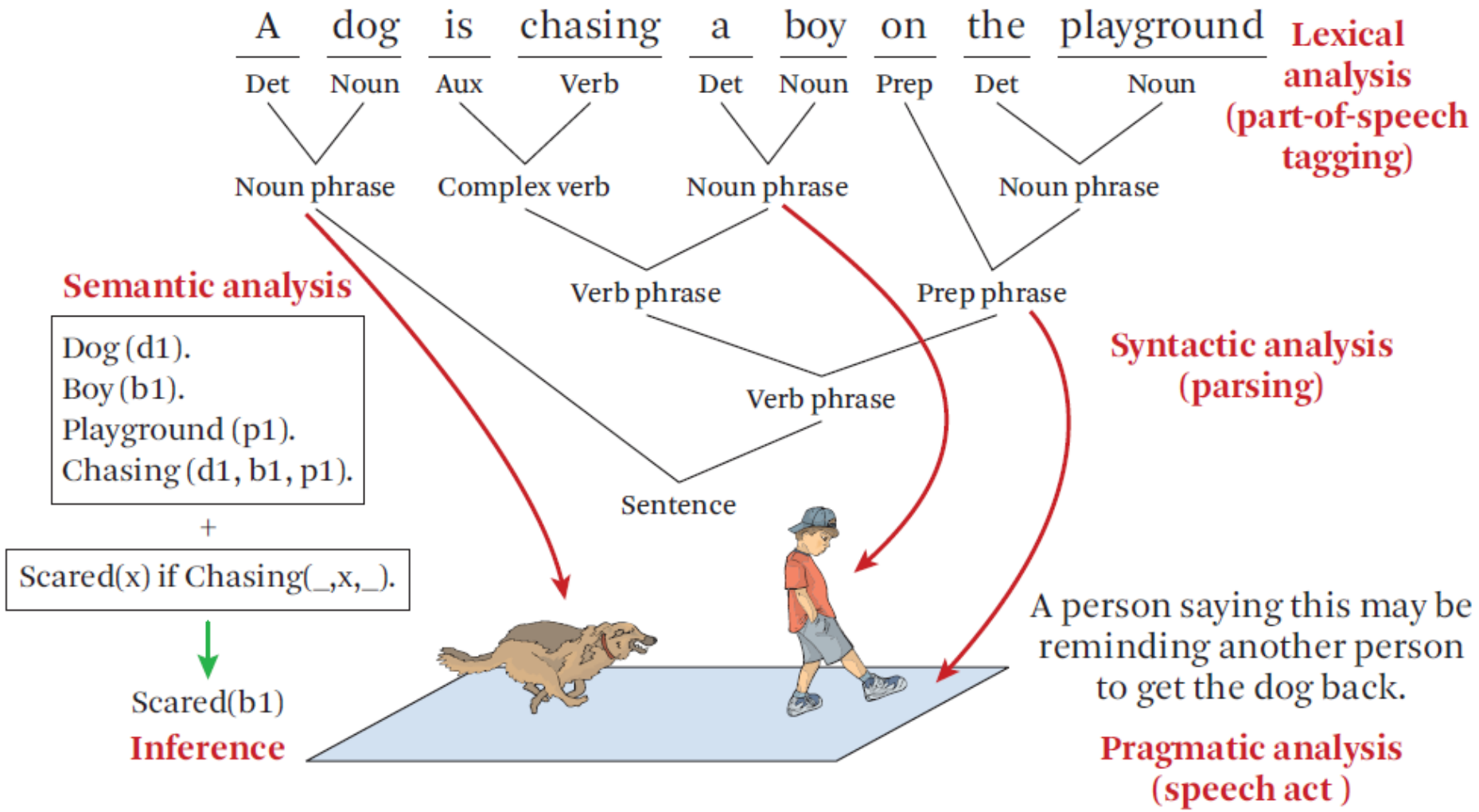


Figure 3.1 An example of tasks in natural language understanding.

- No grammatical rule
- Provides a sequence of tokens such as words, punctuation marks and numbers
- Part of Speech (PoS):
  - The second stage of text processing: each extracted token is assigned a grammatical role: nouns (objects in the sentence), adjectives (describe a noun), adverbs (describe how an action takes place), verbs(show an action), prepositions (define the relative relationship between nouns)

CLASSICAL APPROACHES – TOKENIZING

- Tokenizing and PoS require recognition of **patterns** in a sentence.
- Defines patterns as a sequence of characters.
- Regular expression patterns can be case-sensitive/detect single digits/exclude some characters/detect spelling differences in British and American English/detect alternatives or repetitions
- Applies regular expression patterns by defining hierarchical structures (a top-own path from a sentence to set of words)

CLASSICAL APPROACHES – PRODUCTION RULES FOR LANGUAGE

S → NP + VP	A sentence is composed of a noun phrase and a verb phrase.
NP → PN	A noun phrase can be proper noun.
NP → Det + N	A noun phrase can be a determinant and a noun phrase.
PN → Hans | Paul | Hannah | Stephan | Lara	A proper noun can be either one of the names “Hans”, “Paul”, “Hannah”, “Stephan”, “Lara”.
VP → V	A verb phrase can be a verb.
VP → V + NP	A verb phrase can be a verb and a noun phrase.
Det → the | a | an	A determinant can be a “the” or “a” or “an”.
N → book | student | football |sandwich |bed	A verb can be “book”, “student”, “football”, “sandwich”, or “bed”.
V → studied | played | ate | slept	A verb can be studied, played, ate, or slept.

Source of the table: Course Book DLMBBD01, p. 69

Lexical Analysis

- After word extraction, each word needs to be broken into parts.
- This is necessary for the following stages:
  - **Phonology**
  - **Morphology**

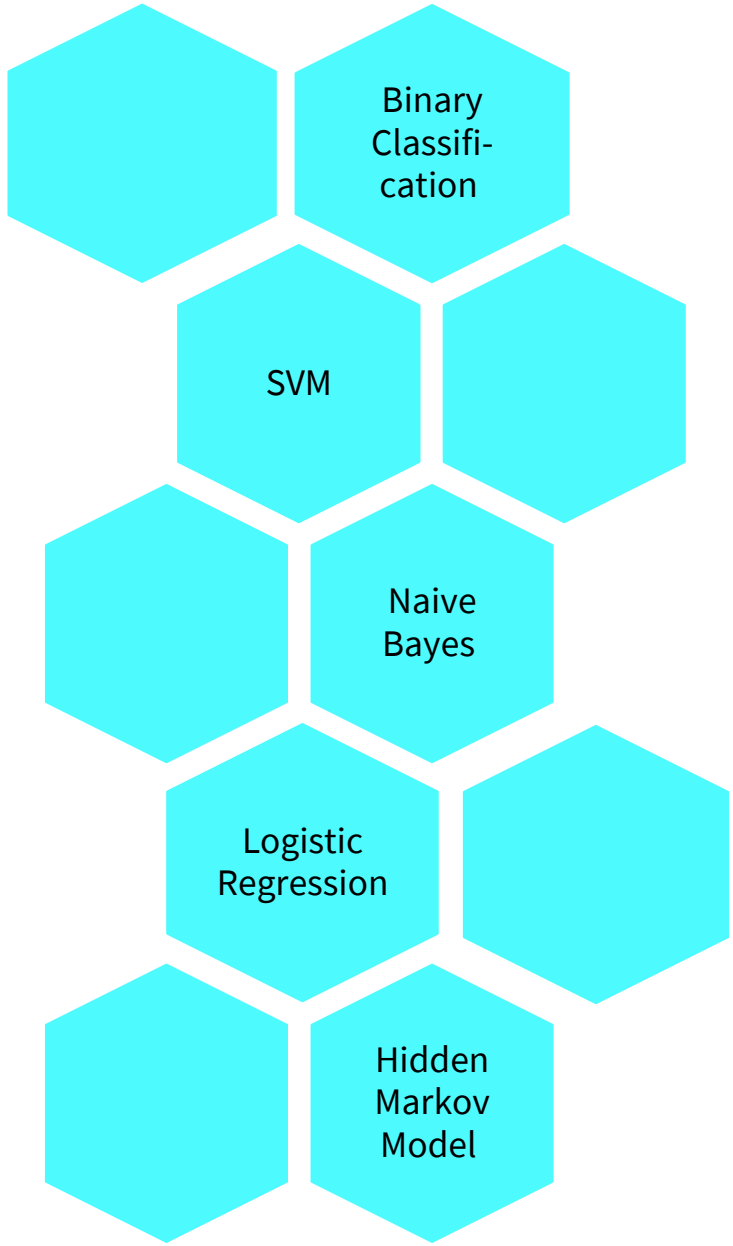
Syntactic Parsing

- A sentence: basic unit of logical analysis, expressing opinions, ideas, or thoughts
- determines syntactic and grammatical structure of sentence (syntax tree, LR parsing, context-free grammar)

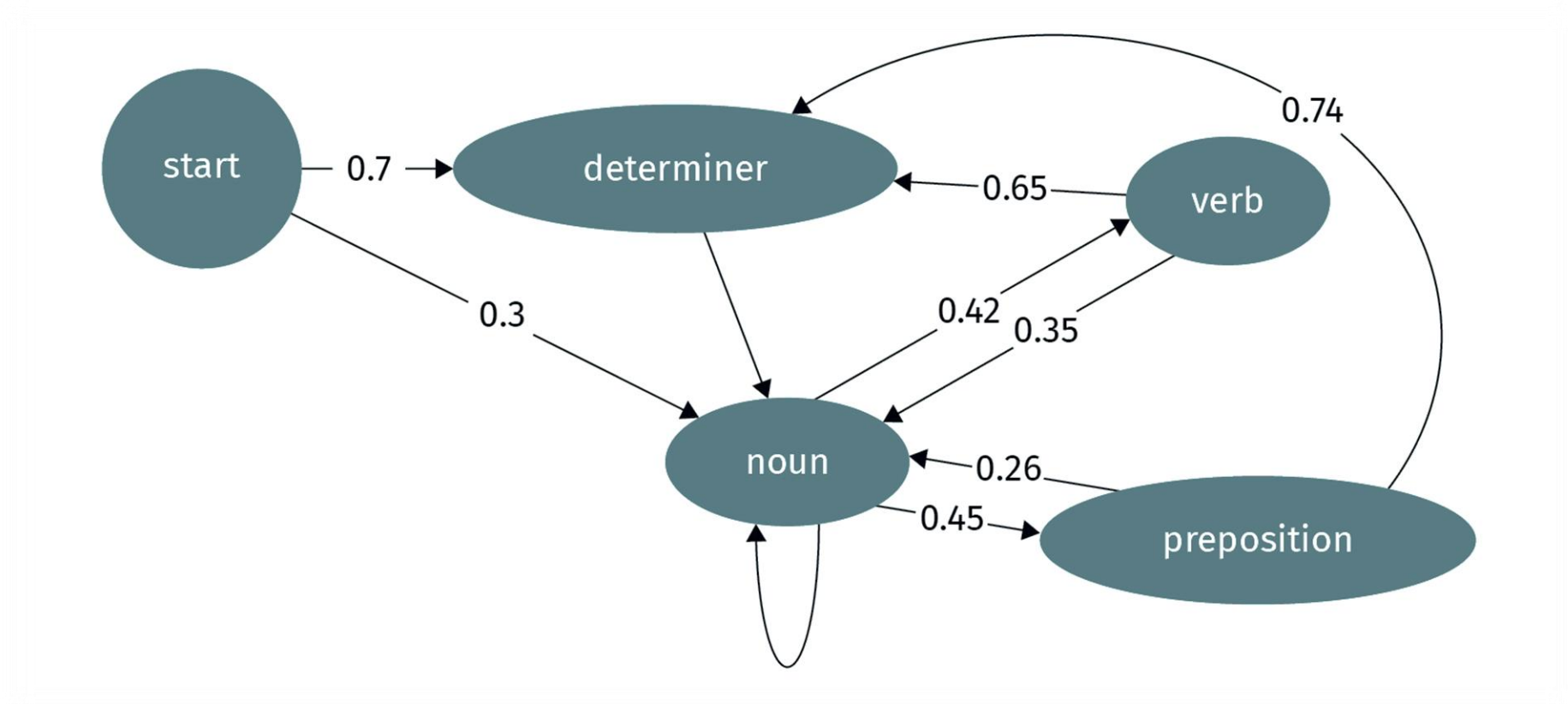
- Analyzing the meaning of words, fixed expressions, whole sentences and utterances in context
- Relating syntactic structures at a phrase, clause or sentence level to the writing as a whole and to their language-independent meaning
  - **Semantics deals with word or sentence choice in any given context (universally-coded meaning).**
  - **Pragmatics considers the unique or particular meaning derived from context (the listerner's interpretation).**

SECOND GENERATION STATISTICAL AND EMPIRICAL APPROACHES

- Supervised learning used to predict POS based on sentences
- Unsupervised techniques used for clustering texts into similar groups
- SVM: discriminative classifier by hyperplane
- hidden Markov model

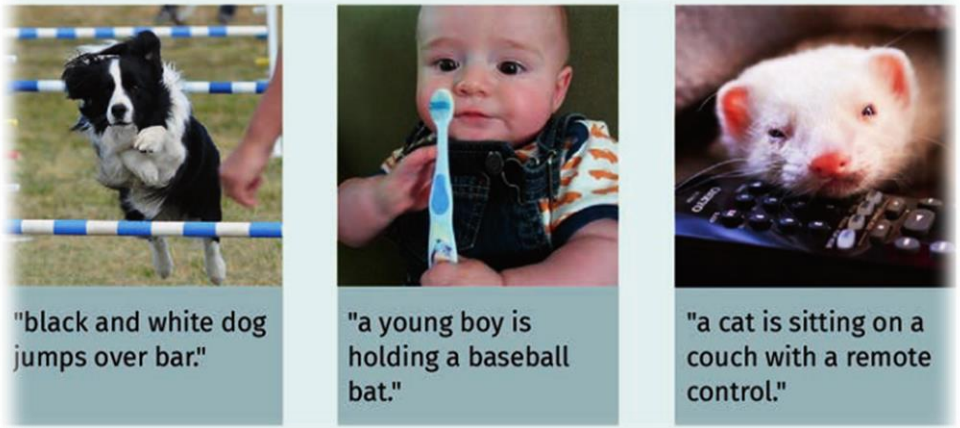


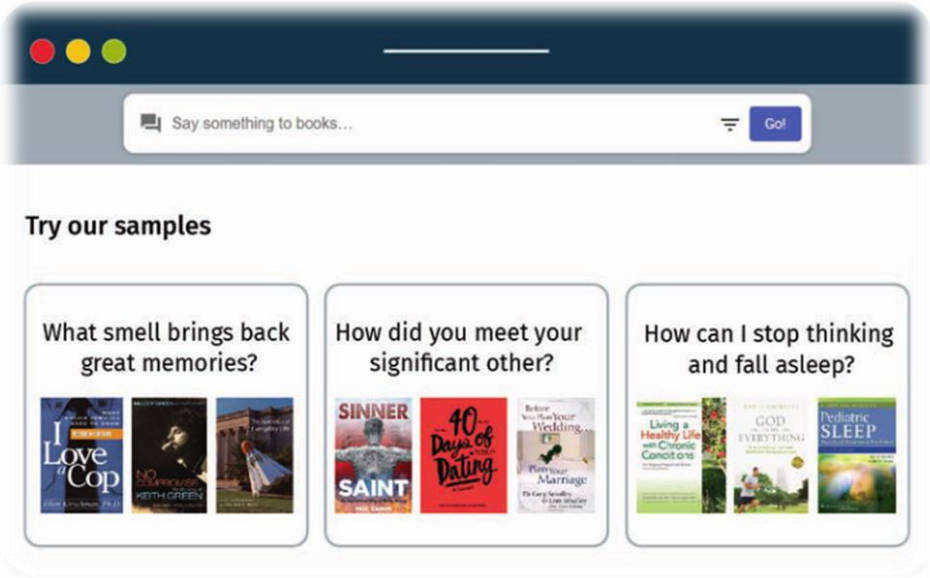
SECOND GENERATION STATISTICAL AND EMPIRICAL APPROACHES



Source of the graphic: Course Book DLMBBD01, p. 72

- RNN is used to generate natural language descriptions of images and their regions.
- **Google Talk to Books** provides a service that uses AI to talk to books and test word association skills - users can converse with a smart algorithm that answers questions by surfacing relevant passages from books.





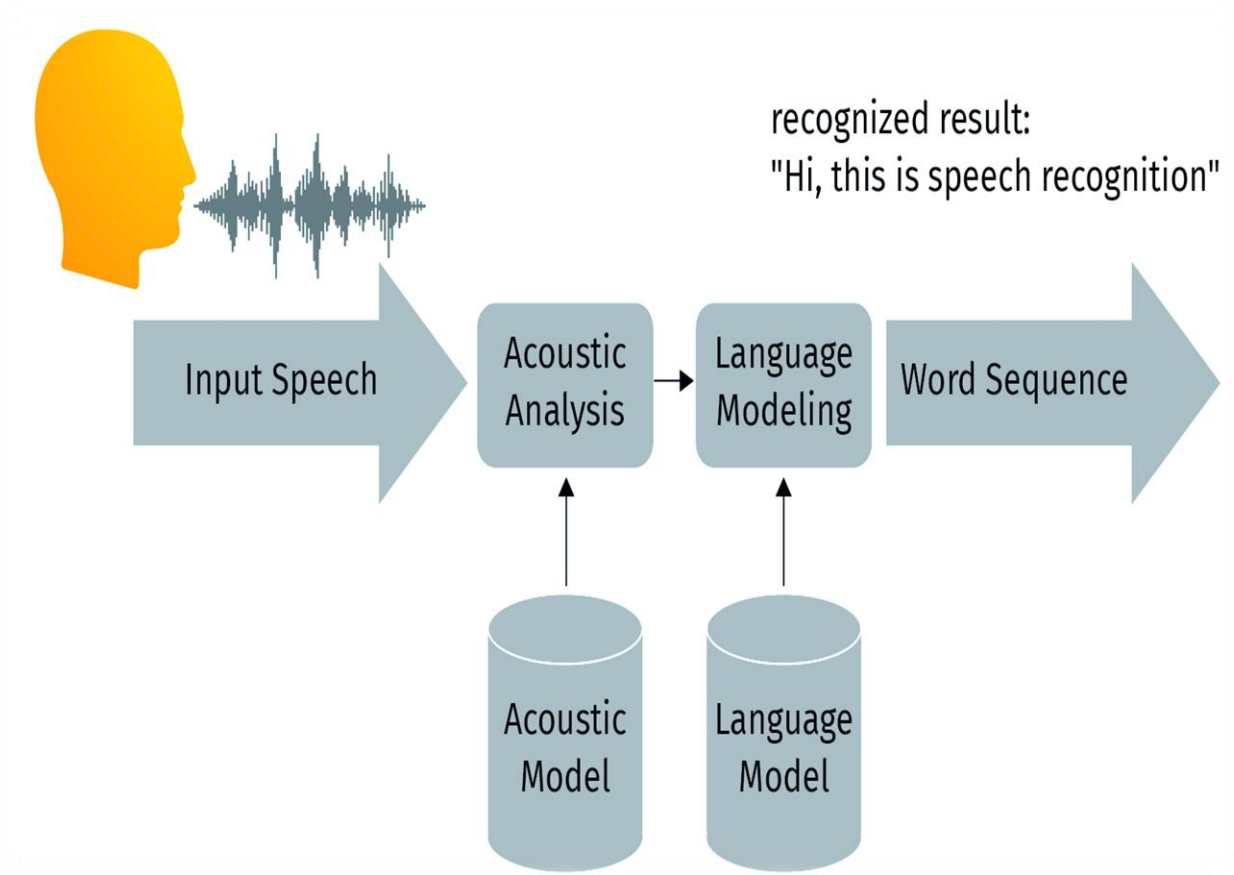
Source of the images: Karpathy & Fei-Fei, 2015; Talk to Books, n.d.

SPEECH RECOGNITION

- Develops methodologies that enables recognition and translation of spoken languages into text by computers
- Known as: computer speech recognition, automated speech recognition(ASR), and speech to text (STT)
- **Applications**: call routing, voice dialing, voice search
- **ASR**: converts an acoustic speech signal into symbolic description of a message coded in the signal during speech production (components: symbolic, grammatical, semantic, pragmatic)

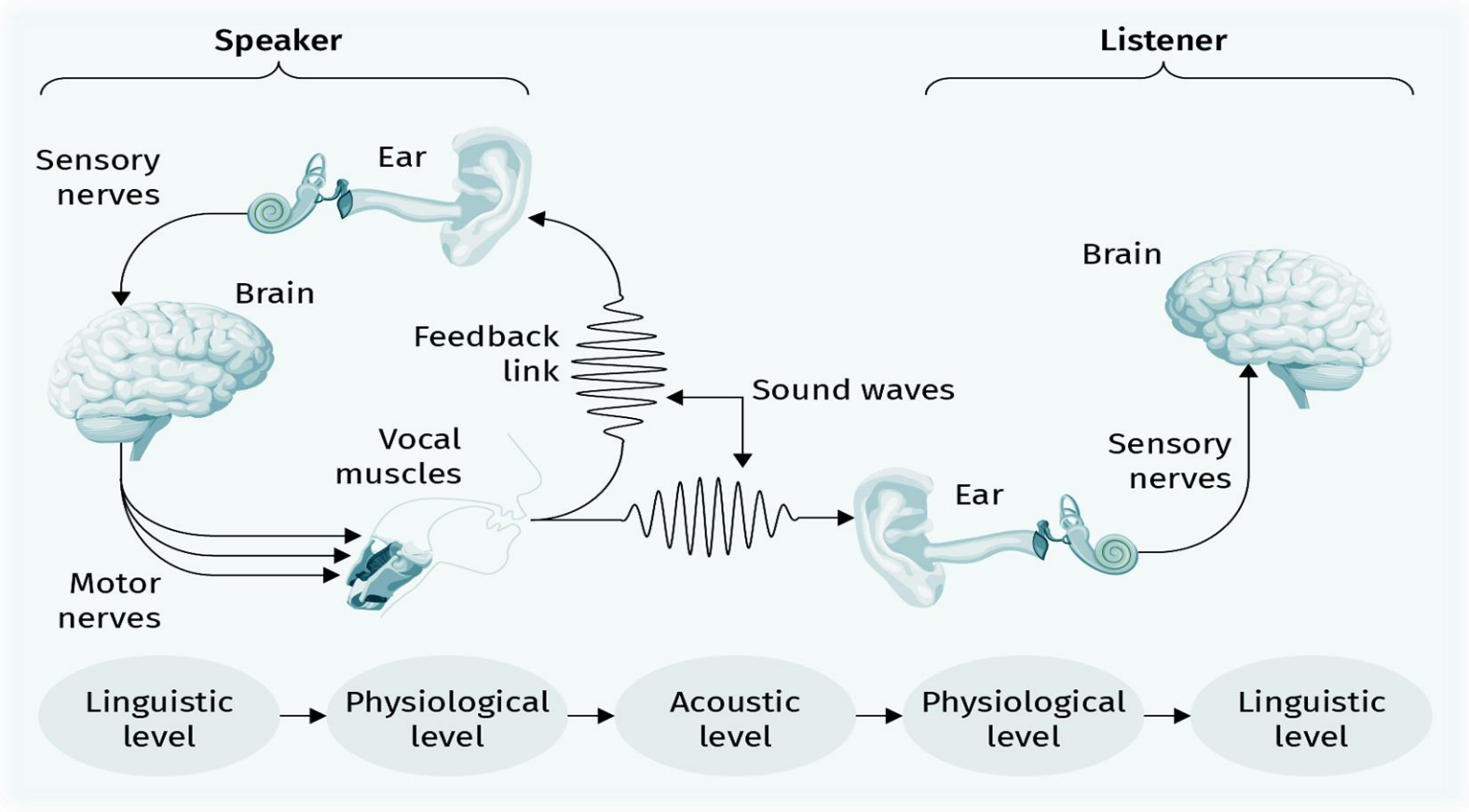
SPEECH RECOGNITION PROCESS

- **Acoustic analysis** represents the intermediary between audio signals and linguistic units of speech.
- **Language analysis** matches sound produced with word sequences to distinguish between familiar words.



Source of the graphic: Course Book DLMBBD01, p. 75

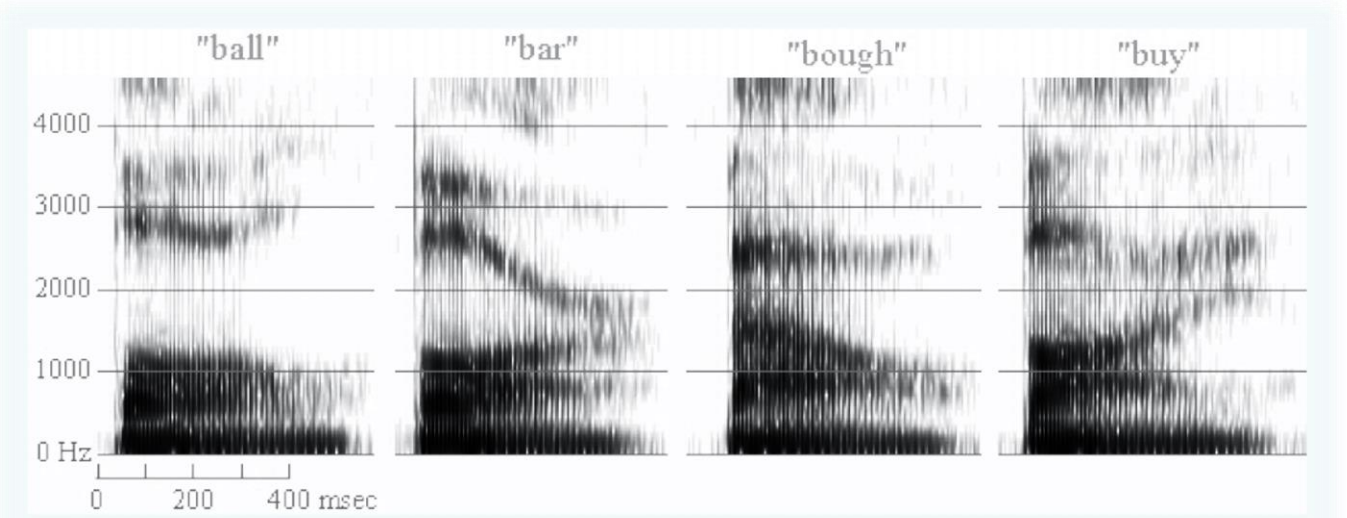
The whole process starts at the linguistic level, proceeds to the physiological and linguistic levels and then returns back to the physiological and linguistic levels of the listener.



Source of the graphic: Denes & Pinson, 1993

SPEECH RECOGNITION MODELS – ACOUSTIC MODEL

- Acoustic model: the engine of the automatic speech recognition (ASR) system
- Represents relationship between an audio signal and the phonemes and other linguistic units making up speech.
- A common way to represent data is a **spectrogram.**



Source of the graphic: Hagiwara, 2009.

SPEECH RECOGNITION MODELS – LANGUAGE MODEL

Language models are developed to resolve ambiguities and distinguish between homophones and phrases that sound similar.

Statistical Language Model

- Estimates likelihood of a given phrase in order to resolve ambiguities / the probability distribution of natural languages over sequences of words estimated.
- apply n-gram method

$$P(w_1 . w_2 . \ldots . w_m) = \prod_{i=1}^{m} P(w_i \mid w_1 . \ldots . w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)} . \ldots . w_{i-1})$$

Neural Network Model

- Neural network are very promising in predicting the sequence of words.
- The network is trained to predict the probability over the vocabulary.

- **Term frequency (TF)** is based on counting the number of times each word appears in the documents of each class:

$$a_{ij} = f_{ij} / \max(f_j)$$

- TF value is normalized by the logarithm of frequency of documents that the word has occurred in:

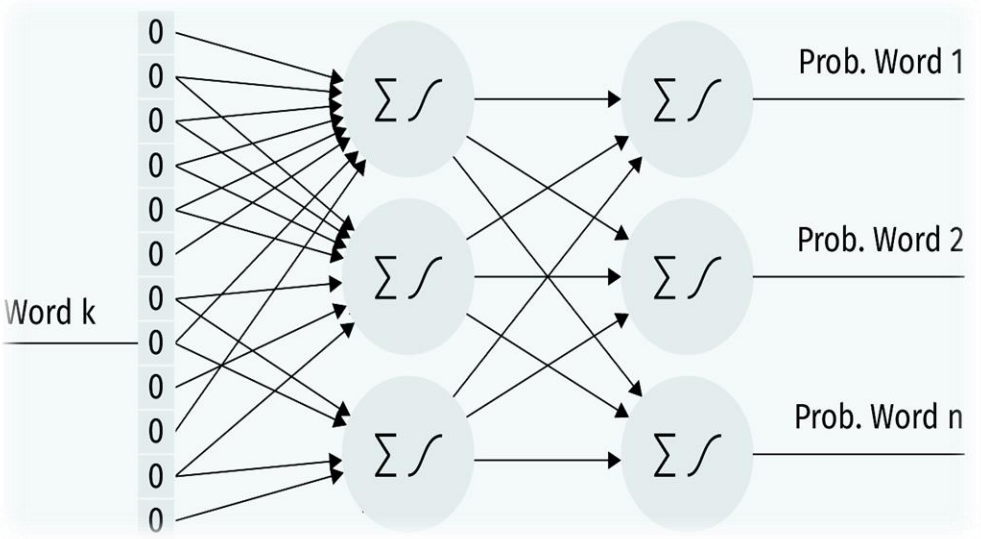
$$t_{ij} = a_{ij} \times \log \frac{N}{df_i}$$

Word Frequency Matrix

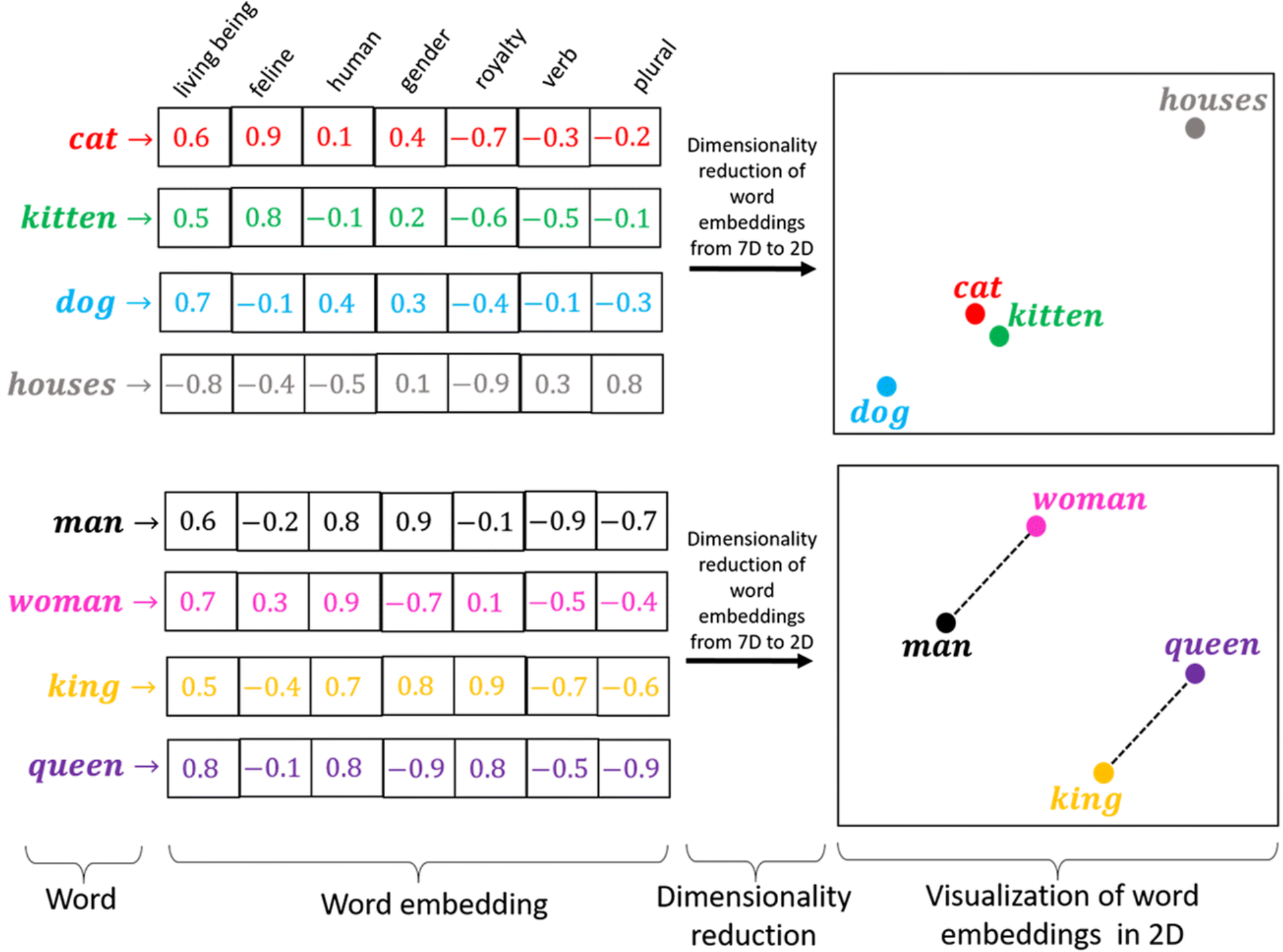
	Word 1	Word 2	Word 3	Word 4	Word 5	Word k	Word n
Document 1	4	10	2	15	…	…	…
Document 2			…	…	…	…	…
Document 3	…	…					
Document 4	…	…	…	…			
Document 5	…	…		…	…	…	
Document k	…	…	…				…
Document n	…	…	…	…	…		…

- Group of methods that evaluates each word based on context; the result is a conditional probability.
- **First method: Continuous bag of words (CBOW)** calculates the probability of observing a word, given a context.
- **Second method: Skip-gram** calculates the probability of observing a specific context given a particular word.

Skip-Gram Word2Vec Diagram



Source of the graphic: Course Book DLMBBD01, p. 80


living being
feline
human
gender
royalty
verb
plural
cat → 0.6 0.9 0.1 0.4 −0.7 −0.3 −0.2
kitten → 0.5 0.8 −0.1 0.2 −0.6 −0.5 −0.1
dog → 0.7 −0.1 0.4 0.3 −0.4 −0.1 −0.3
houses → −0.8 −0.4 −0.5 0.1 −0.9 0.3 0.8
Dimensionality reduction of word embeddings from 7D to 2D
houses
cat
kitten
dog
man → 0.6 −0.2 0.8 0.9 −0.1 −0.9 −0.7
woman → 0.7 0.3 0.9 −0.7 0.1 −0.5 −0.4
king → 0.5 −0.4 0.7 0.8 0.9 −0.7 −0.6
queen → 0.8 −0.1 0.8 −0.9 0.8 −0.5 −0.9
Dimensionality reduction of word embeddings from 7D to 2D
woman
man
queen
king
Word
Word embedding
Dimensionality reduction
Visualization of word embeddings in 2D

Thou shalt not make a machine in the likeness of a human mind

Sliding window across running text

thou	shalt	not	make	a	machine	in	the	…
thou	shalt	not	make	a	machine	in	the	
thou	shalt	not	make	a	machine	in	the	
thou	shalt	not	make	a	machine	in	the	
thou	shalt	not	make	a	machine	in	the	

Dataset

input 1	input 2	output
thou	shalt	not
shalt	not	make
not	make	a
make	a	machine
a	machine	in

SPEECH RECOGNITION APPLICATION

VOICE ASSISTANT

- Microsoft Cortana, Google Assistant, Apple's Siri and Amazon Alexa
- Apply speech recognition technique and AI to comprehend user's voice, analyze it and create a proper response.

EDUCATION

- Used for educational purposes especially in learning a second language.
- Students learn right pronunciation and improve their speaking level.

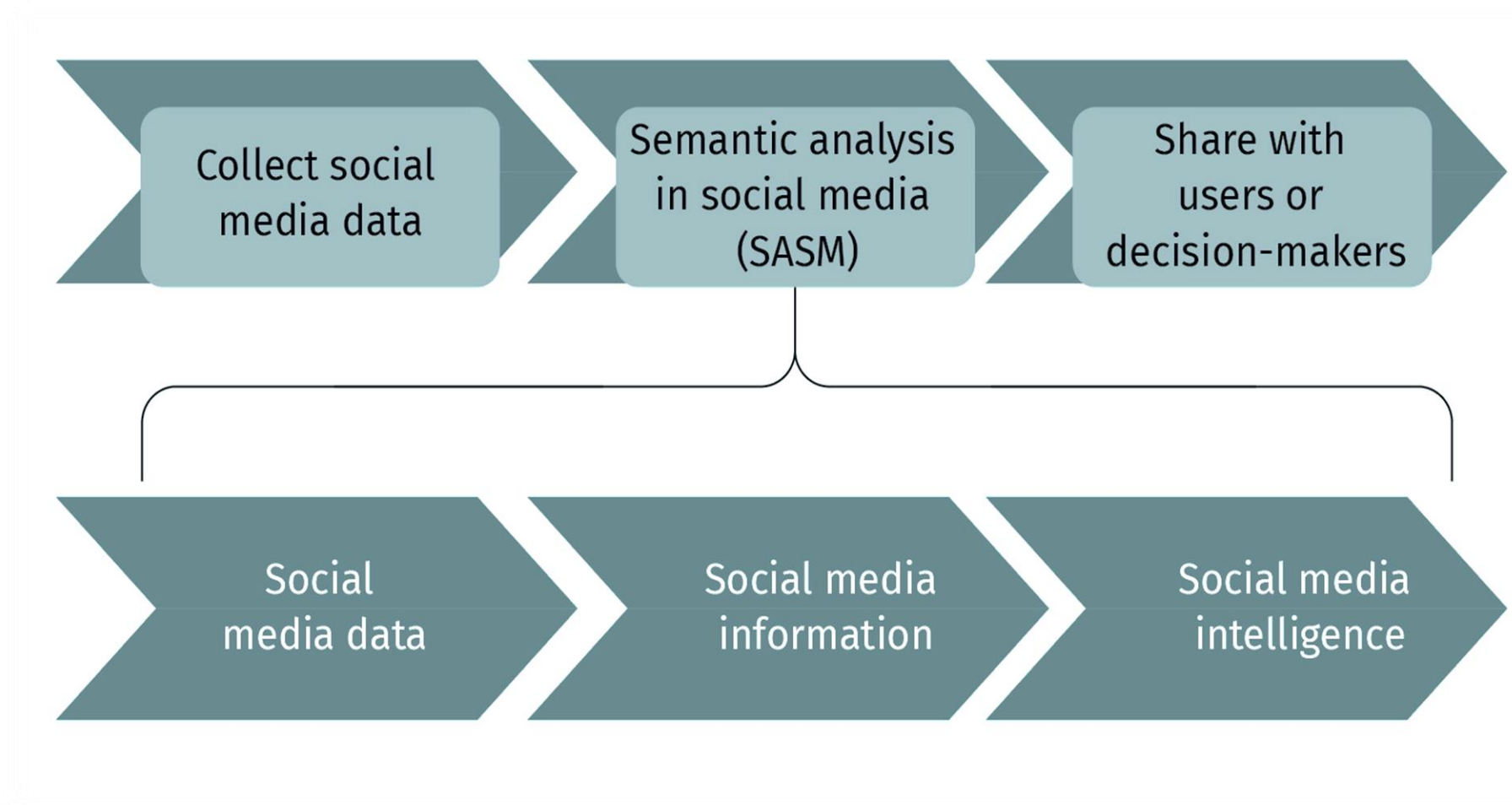
SMART DEVICES

- Equipped with voice recognition - receive commands from users which they recognize and then execute.

PEOPLE WITH DISABILITIES

- Blind people benefit from systems capable of receiving voice commands or that convert Web content into sound.

SOCIAL MEDIA ANALYTICS – FRAMEWORK FOR SEMANTIC ANALYSIS IN SOCIAL MEDIA



Source of the graphic: Farzindar & Inkpen, 2015.

NLP, LINGUISTIC PROCESSING METHODS AND TOOLS USED IN SOCIAL MEDIA ANALYSIS

TOKENIZER

- A tool which separates words from punctuations and other symbols.
- Attention: Punctuation often indicates the end of the phrase, sometimes appears in the abbreviated form of the words.

PART-OF-SPEECH TAGGER

- Labels each word according to their POS.
- Needs pre-defined tagged data to train the model.

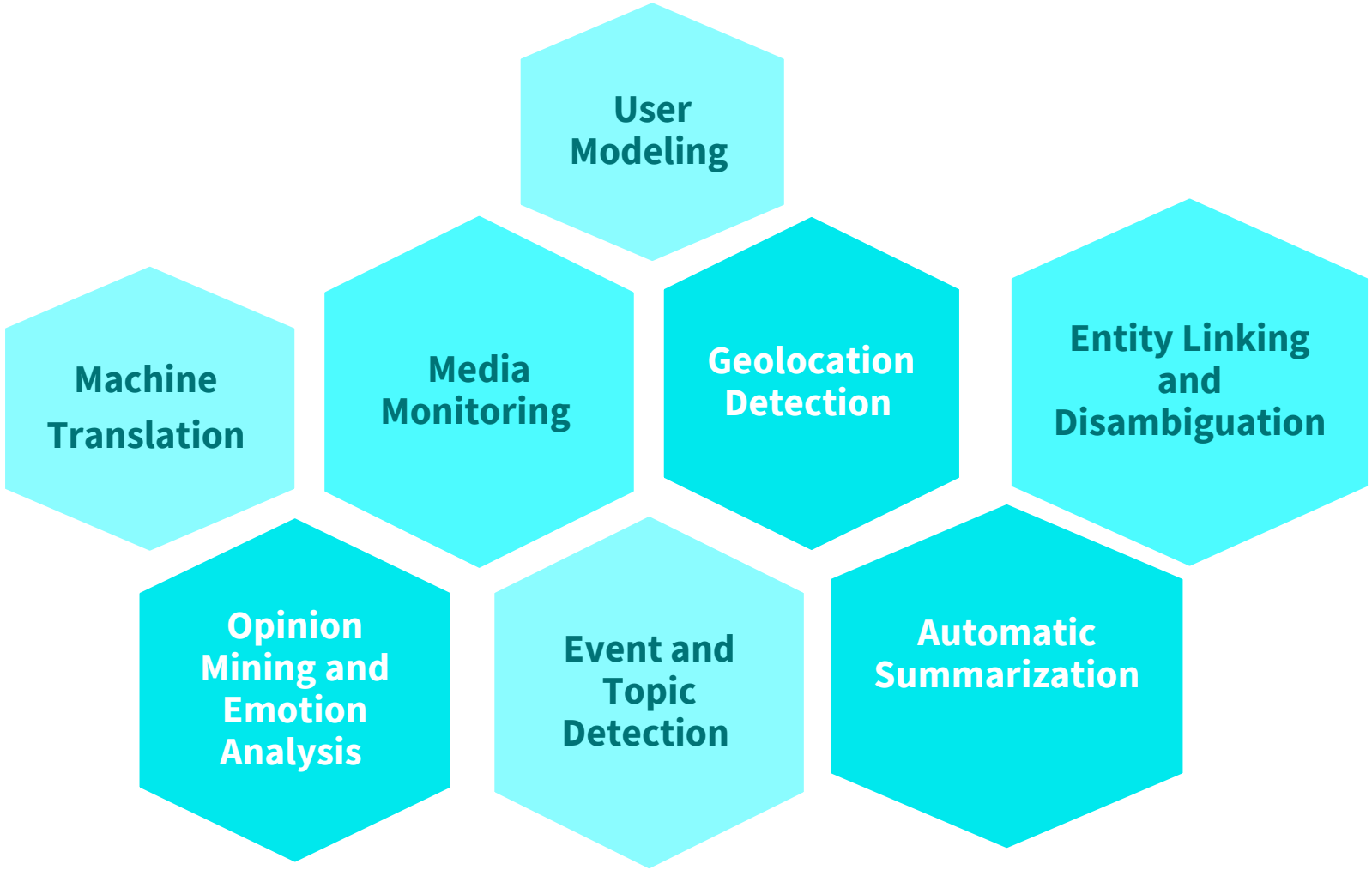
NAMED ENTITY RECOGNIZER

- Semantic analysis : detecting entity and determining the type

CHUNKER AND PARSERS

- Chunker detects the elementary component of a sentence.
- Parser(syntactic analysis): building a parse tree used in further processing such as semantics
- Dependency parser: extracts pair of words that are in a syntactic dependency relationship

SOCIAL MEDIA ANALYTICS APPLICATION



On completion of this unit, you will have learned ...

- … concepts of natural languages.
- … the three generations of natural language processing (NLP).
- … underlying concepts and models of speech recognition.
- … applications of NLP in social media analysis.
- … concepts of and techniques for different types of analyses, such as lexical, syntactical,and semantic analysis.

SESSION 3

TRANSFER TASK

If you want to summarize a text with machine learning and NLP algorithms, how would you do that ? Can NLP summarize a text close to human expression?

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.

The results will be discussed in plenary.

1. What are the layers (phases) of classical NLP approaches?

   a) tokenization, lexical analysis, syntactic analysis, semantic analysis, pragmatic analysis
   b) tokenization, syntactic analysis, semantic analysis
   c) acoustic analysis, syntactic analysis, semantic analysis, pragmatic analysis
   d) phonem analysis, tokenization, syntactic analysis, semantic analysis

2. Which item is not an application of NLP in social media?

   a) sentiment analysis
   b) automatic summarization
   c) lexical analysis
   d) geo-localization

3. A spectogram is a common way to represent…

a) … language data.
b) … grammatical data.
c) … semantic data.
d) … acoustic data.

LIST OF SOURCES

Araki, Y., Zeng, M. & Huganir R. (2015). Rapid dispersion of SynGAP from synaptic spines triggers AMPA receptor insertion and spine enlargement during LTP. *Neuron 85*(1). 173-189.

Atefeh, F. & Inkpen, D. (2015). Natural language processing for social media. *Synthesis Lectures on Human Language Technologies 8*(2). 1-166.

Denes, P. B. & Pinson, E. N. (1993). *The Speech Chain: The Physics and Biology of Spoken Language*. W. H. Freeman.

Farzindar, A. & Inkpen, D. (2015). *Natural language processing for social media.* Morgan & Claypool.

Hagiwara, R. (2009). *Monthly mystery spectrogram webzone* [guide]. https://home.cc.umanitoba.ca/~robh/howto.html

Indurkhya, N. & Damerau, F. J. (2010). *Handbook of natural language processing* (2nd ed.). CRC Press.

Karpathy, A. & Li, F.-F. (2015). Deep visual-semantic alignments for generating image descriptions.. *CVPR* (p./pp. 3128-3137), : IEEE Computer Society. ISBN: 978-1-4673-6964-0

Talk to Books. (n.d.). Website. https://books.google.com/talktobooks/

Whiteside, Sandra P. "Peter B. Denes and Elliot N. Pinson The Speech Chain: The Physics and Biology of Spoken Language, Oxford: WH Freeman and Company, 1993. Pp. 246 ISBN 0-7167-2344-1." *Journal of the International Phonetic Association 23*(2). 98-101.