




**ĐẠI HỌC ĐÀ NẴNG**  
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG VIỆT - HÀN**  
**Vietnam - Korea University of Information and Communication Technology**

# Natural Language Processing
















## NLP

<http://vku.udn.vn/>

1



## NLP tasks and applications

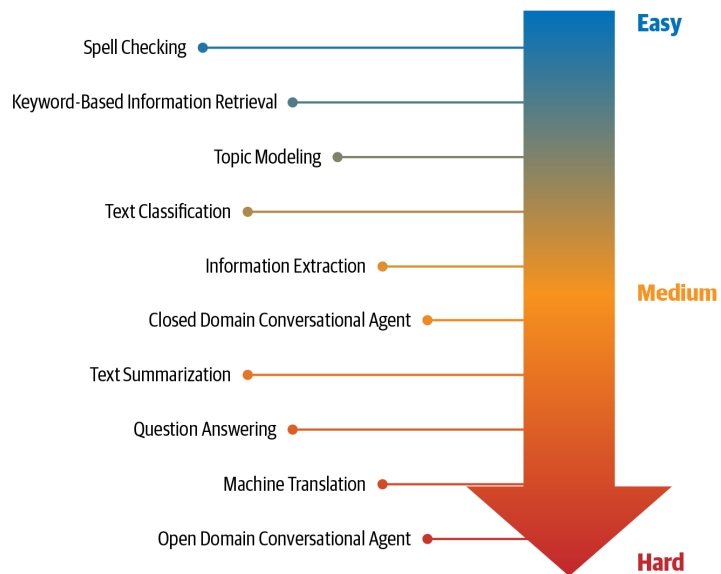
<b>Core Tasks</b> <i>Covered in Chapters 3-7</i>	 Text Classification	 Information Extraction	 Conversational Agent	 Information Retrieval	 Question Answering Systems
<b>General Applications</b> <i>Covered in Chapters 4-7</i>	 Spam Classification	 Calendar Event Extraction	 Personal Assistants	 Search Engines	 Jeopardy!
<b>Industry Specific</b> <i>Covered in Chapters 8-10</i>	 Social Media Analysis	 Retail Catalog Extraction	 Health Records Analysis	 Financial Analysis	 Legal Entity Extraction

<http://vku.udn.vn/>

2

## NLP tasks

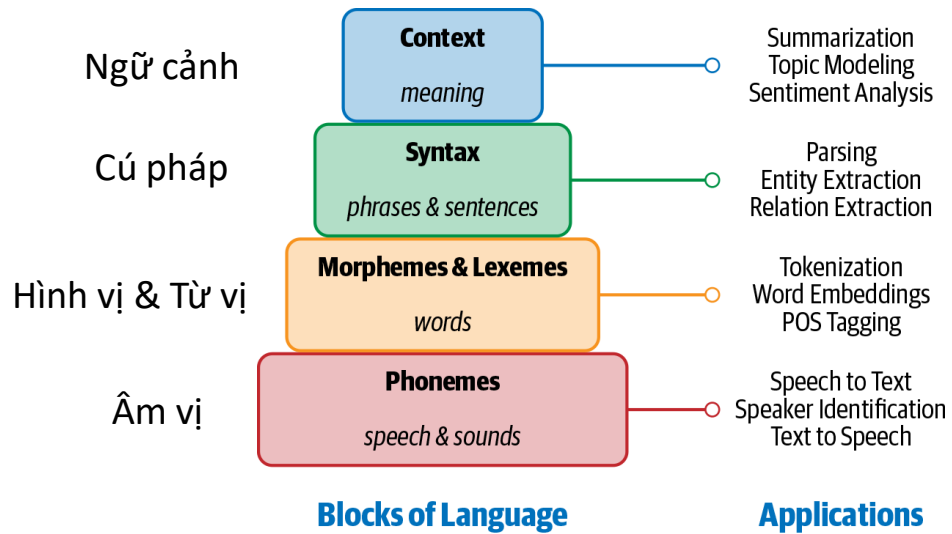
- NLP tasks organized according to their relative difficulty



3



## Building Blocks of Language



<http://vku.udn.vn/>

4



## Phonemes

- Phonemes are the smallest units of sound in a language.
- They may not have any meaning by themselves but can induce meanings when uttered in combination with other phonemes.

Consonant phonemes, with sample words		Vowel phonemes, with sample words	
1. /b/ - bat	13. /s/ - sun	1. /a/ - ant	13. /oi/ - coin
2. /k/ - cat	14. /t/ - tap	2. /e/ - egg	14. /ar/ - farm
3. /d/ - dog	15. /v/ - van	3. /i/ - in	15. /or/ - for
4. /f/ - fan	16. /w/ - wig	4. /o/ - on	16. /ur/ - hurt
5. /g/ - go	17. /y/ - yes	5. /u/ - up	17. /air/ - fair
6. /h/ - hen	18. /z/ - zip	6. /ai/ - rain	18. /ear/ - dear
7. /j/ - jet	19. /sh/ - shop	7. /ee/ - feet	19. /ure/ <sup>4</sup> - sure
8. /l/ - leg	20. /ch/ - chip	8. /igh/ - night	20. /ə/ - corner (the 'schwa' - an unstressed vowel sound which is close to /u/)
9. /m/ - map	21. /th/ - thin	9. /oa/ - boat	
10. /n/ - net	22. /th/ - then	10. /oo/ - boot	
11. /p/ - pen	23. /ng/ - ring	11. /oo/ - look	
12. /r/ - rat	24. /zh/ <sup>3</sup> - vision	12. /ow/ - cow	

<http://vku.udn.vn/>

5



## MORPHEMES AND LEXEMES

- A morpheme is the smallest unit of language that has a meaning.
- It is formed by a combination of phonemes.
- Not all morphemes are words, but all prefixes and suffixes are morphemes.

unbreakable  
*un + break + able*

cats  
*cat + s*

tumbling  
*tumble + ing*

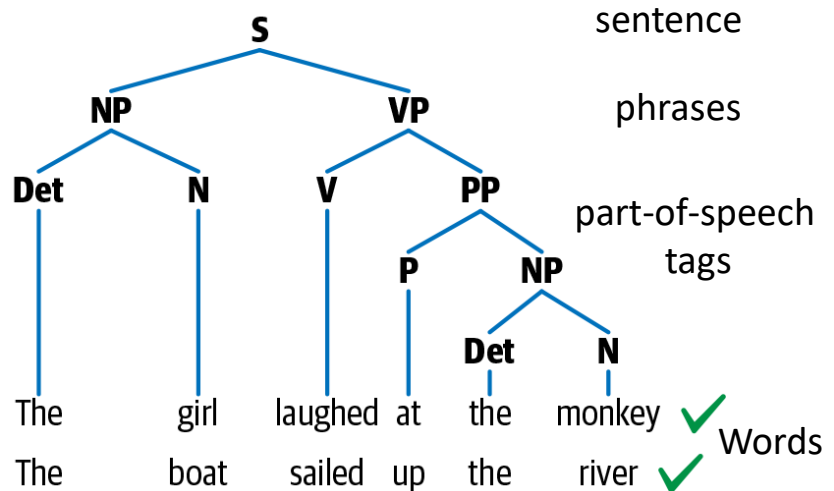
unreliability  
*un + rely + able + ity*

<http://vku.udn.vn/>

6

# SYNTAX

- Syntax is a **set of rules** to construct grammatically correct sentences out of words and phrases in a language.



- N stands for noun, V for verb, and P for preposition. Noun phrase is denoted by NP and verb phrase by VP

7

## VKU CONTEXT

- Context is how various parts in a language come together to convey a particular meaning.
- The meaning of a sentence can change based on the context, as words and phrases can sometimes have multiple meanings.

<http://vku.udn.vn/>

8



## Why Is NLP Challenging?

- **AMBIGUITY**
  - Ambiguity means uncertainty of meaning
- **COMMON KNOWLEDGE**
  - It is the set of all facts that most humans are aware of.
  - For example
    - “man bit dog”
    - “dog bit man.”
- **CREATIVITY**
- **DIVERSITY ACROSS LANGUAGES**

<http://vku.udn.vn/>

9



## Bài tập

- Chọn 1 bài toán XLNNTN và tìm hiểu:
  - Bài toán xử lý vấn đề gì?
  - Tổng quan về các giải pháp xử lý
- Sử dụng một trong các thư viện NLP và tìm hiểu câu lệnh xử lý:
  - Stopwords
  - Similar words
  - word\_tokenize
  - sentence\_tokenize
  - Stemming and Lemmatization: phân biệt? câu lệnh?
  - POST Tagging: khái niệm? câu lệnh? (Tùy chọn)
- Tìm hiểu thư viện **underthesea**

<http://vku.udn.vn/>

10