

An illustration of a computer lab or classroom. Several students are seated at desks, each with a computer monitor and keyboard. They are all wearing large headphones, suggesting they are practicing listening or speaking exercises. The scene is dimly lit, with the primary light source being the screens of the computers. The students are focused on their work, with some looking at the screens and others at open books or papers.

GIỚI THIỆU HỌC PHẦN

XỬ LÝ NGÔN NGỮ TỰ NHIÊN

GIỚI THIỆU HỌC PHẦN



- ❖ Nội dung học phần
- ❖ Thời lượng: 30 tiết lý thuyết + 30 tiết thực hành
- ❖ Mục tiêu học phần
- ❖ Đánh giá học phần
 - Giữa kỳ
 - Cuối kỳ



Chương 1

TỔNG QUAN

XỬ LÝ NGÔN NGỮ TỰ NHIÊN

Add a footer

Ngôn ngữ tự nhiên (Natural language)

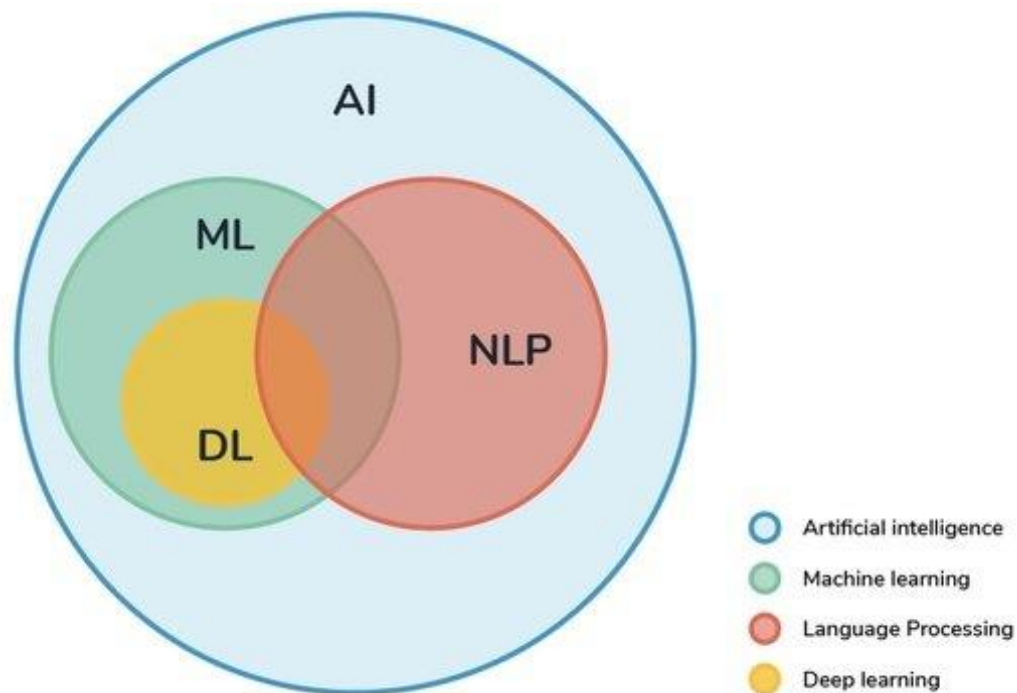
- ❖ **Dữ liệu dạng văn bản (textual data):** dữ liệu phi cấu trúc
 - Thuộc về một ngôn ngữ
 - Có ngữ nghĩa và cú pháp
- ❖ **Ngôn ngữ tự nhiên (natural language):** ngôn ngữ được phát triển, tiến hóa và sử dụng bởi con người
 - Hệ thống chữ viết
 - Hệ thống tiếng nói
 - Hệ thống ký hiệu

Xử lý ngôn ngữ tự nhiên (Natural Language Processing)

❖ Xử lý ngôn ngữ tự nhiên (Natural Language Processing — NLP)

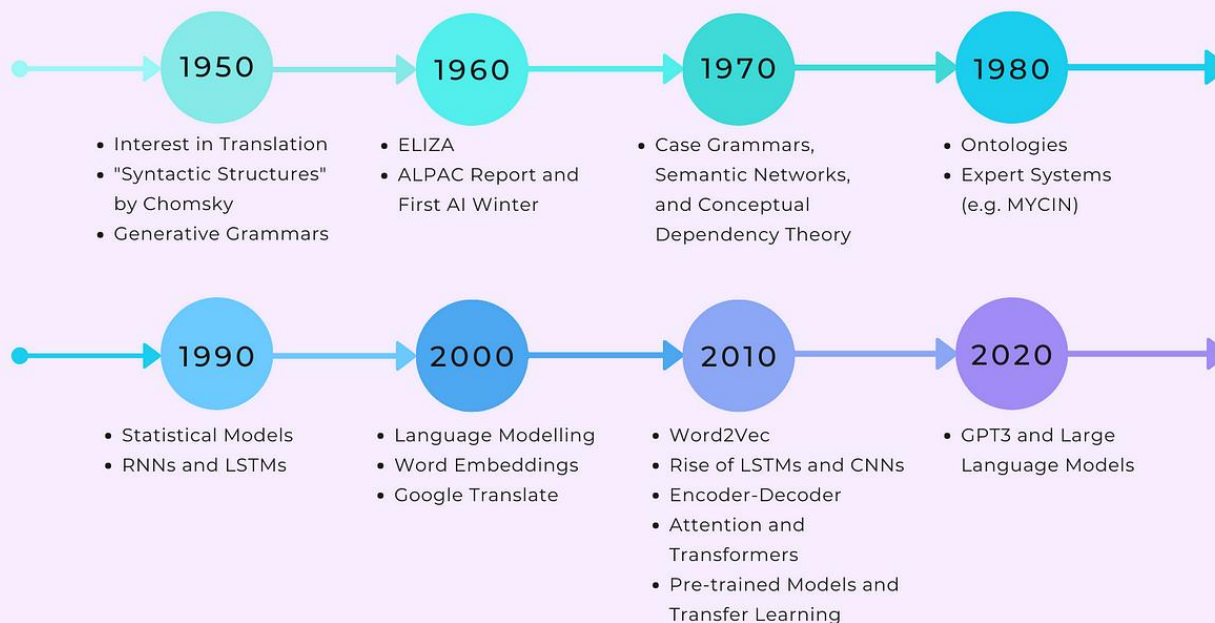
- Một nhánh của trí tuệ nhân tạo cho phép máy tính đọc, trích xuất ngữ nghĩa từ văn bản và tạo ra tài liệu (document)
- Nghiên cứu, phân tích nhằm hiểu và tạo ngôn ngữ để giao tiếp với con người cả nói và viết
- Goldberg (2017): NLP là lĩnh vực **phân tích, thiết kế** thuật toán với input và output là dữ liệu không có cấu trúc hoặc ngôn ngữ tự nhiên.
- Eisenstein (2018): NLP tập trung **phân tích thiết kế** các thuật toán và cách thức biểu diễn để xử lý ngôn ngữ tự nhiên của con người.

Xử lý ngôn ngữ tự nhiên (Natural Language Processing)



Lịch sử phát triển

A Brief Timeline of NLP



Tại sao NLP là lĩnh vực đầy thách thức?

❖ Ngôn ngữ **đa nghĩa (ambiguous)**

- Các đơn vị ngôn ngữ (từ, cụm từ) có thể mang nhiều nghĩa

❖ Ngôn ngữ có tính **hợp thành (compositional)**

- Ý nghĩa của một đơn vị ngôn ngữ được xác định bởi ý nghĩa của các thành phần cấu tạo nên nó

❖ Ngôn ngữ có tính **đệ quy (recursive)**

- Các đơn vị ngôn ngữ có thể được kết hợp lặp đi lặp lại để tạo thành các cấu trúc phức tạp hơn

❖ Ngôn ngữ ẩn chứa **cấu trúc (hidden structure)**

- Những thay đổi nhỏ trong một câu có thể ảnh hưởng đến toàn bộ ý nghĩa của câu

Xử lý ngôn ngữ tự nhiên (Natural Language Processing)

❖ Một số ứng dụng NLP:

- chatbot: ChatGPT, Bing Chat
- trợ lý ảo (virtual assistant): Siri, Alexa, Google Home
- dịch máy (machine translation): Google Translate, DeepL
- phân tích cảm xúc (sentiment analysis)
- phát hiện tin giả (fake news detection)



Xử lý ngôn ngữ tự nhiên (Natural Language Processing)

❖ NLP là cơ sở của các ứng dụng AI tạo sinh (Generative AI)

- AlphaCode / Github Copilot (text → code)
- DALL-E / Midjourney (text → image)
- Pika / Lumiere / Sora (text → video)



Ứng dụng của NLP

❖ Kiểm tra chính tả và ngữ pháp

➤ Gợi ý sửa lỗi

The screenshot shows a Google search for "grammar check". The search bar at the top contains the text "grammar check" with a search icon on the right. Below the search bar, there are tabs for "Tất cả", "Hình ảnh", "Video", "Mua sắm", "Web", "Tin tức", "Sách", and "Thêm", with "Tất cả" selected. The search results are displayed below the tabs. The first result is for "Grammarly", with the URL "https://www.grammarly.com/grammar-checker". The second result is for "GrammarCheck", with the URL "https://www.grammarcheck.net/editor". The advertisement on the right side of the page is for "Grammarly" and features a screenshot of the Grammarly interface. The advertisement text reads: "Công cụ kiểm tra ngữ pháp (Grammar checker)". The screenshot shows a text editor with the sentence "community is in need of." and a suggestion to change it to "community needs". Below the screenshot, the text reads: "Trình kiểm tra ngữ pháp, theo thuật ngữ máy tính, là một chương trình hoặc một phần của chương trình, có gắng xác minh văn bản viết về tính đúng ngữ pháp. Wikipedia (tiếng Anh) >".

Google

grammar check

Tất cả Hình ảnh Video Mua sắm Web Tin tức Sách Thêm Công cụ

Đang hiển thị kết quả cho **grammar check**
Tìm kiếm thay thế cho **grammar check**

Được tài trợ

Grammarly
<https://www.grammarly.com/grammar-checker>

Grammar Check
Check Your Grammar in Seconds — With just a few clicks, clean up typos, grammatical mistakes, and misplaced punctuation. Eliminate **grammar** errors instantly and enhance your writing. Try it now for free!

GrammarCheck
<https://www.grammarcheck.net/editor> · Dịch trang này

Free Grammar Checker (Online Editor)
No sign-up required ✓ It's simple: copy and paste your text into the online editor to check grammar, spelling, and punctuation. Find the best words to ...
[Infographics · Blog · 16 Boring Words · Fish or Fishes? 16 Tricky...](#)

Công cụ kiểm tra ngữ pháp (Grammar checker)

community **is in need of.**
→ **needs**
The phrase **is in need of** may be wordy. Consider changing the wording.

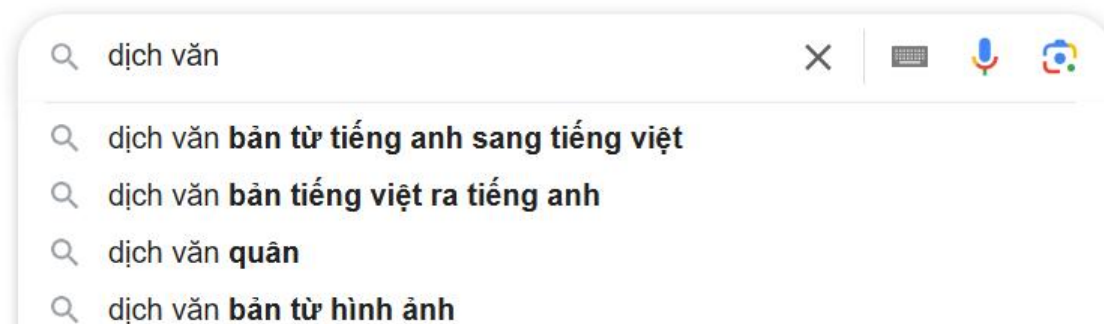
GrammarCheck
167

Trình kiểm tra ngữ pháp, theo thuật ngữ máy tính, là một chương trình hoặc một phần của chương trình, có gắng xác minh văn bản viết về tính đúng ngữ pháp.
[Wikipedia \(tiếng Anh\) >](#)

Ứng dụng của NLP

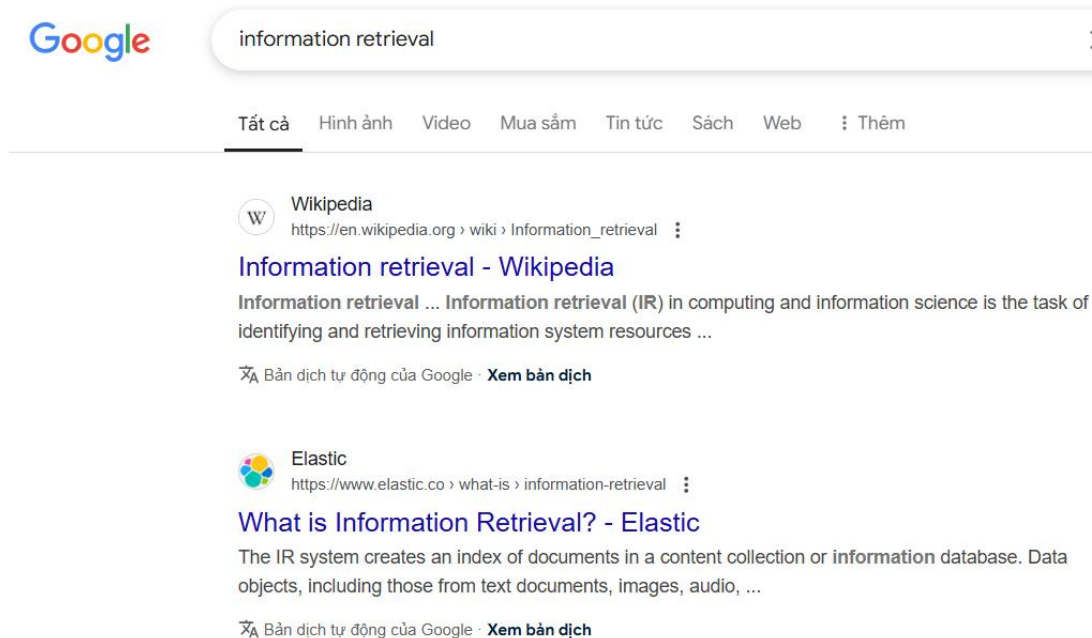
❖ Word prediction

- Dự đoán từ tiếp theo người dùng có khả năng nhập vào



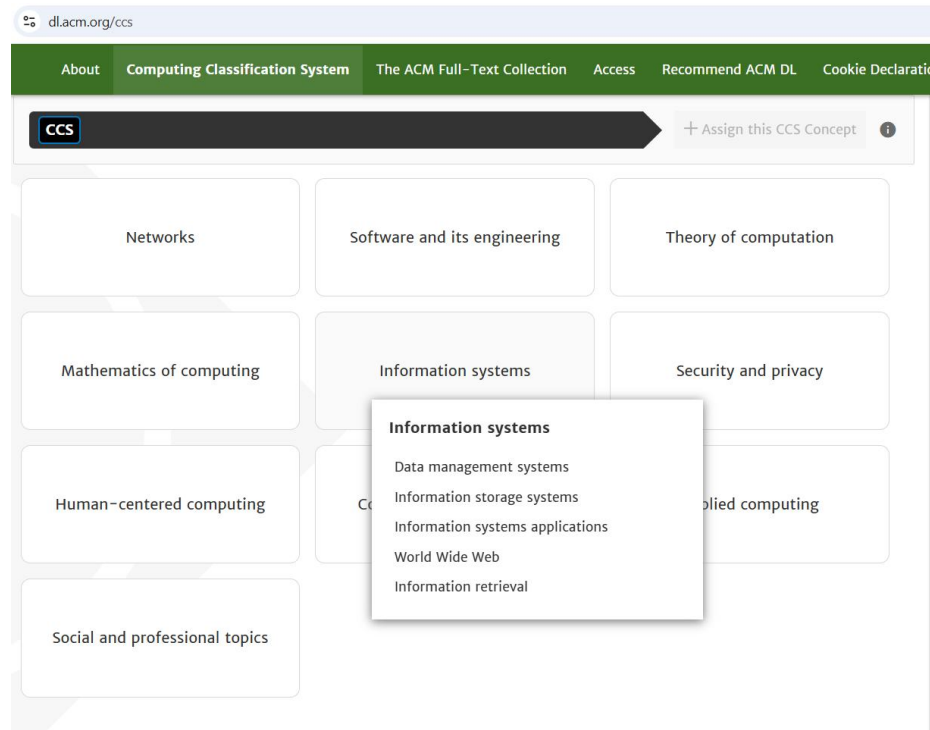
Ứng dụng của NLP

❖ Truy vấn thông tin (information retrieval)



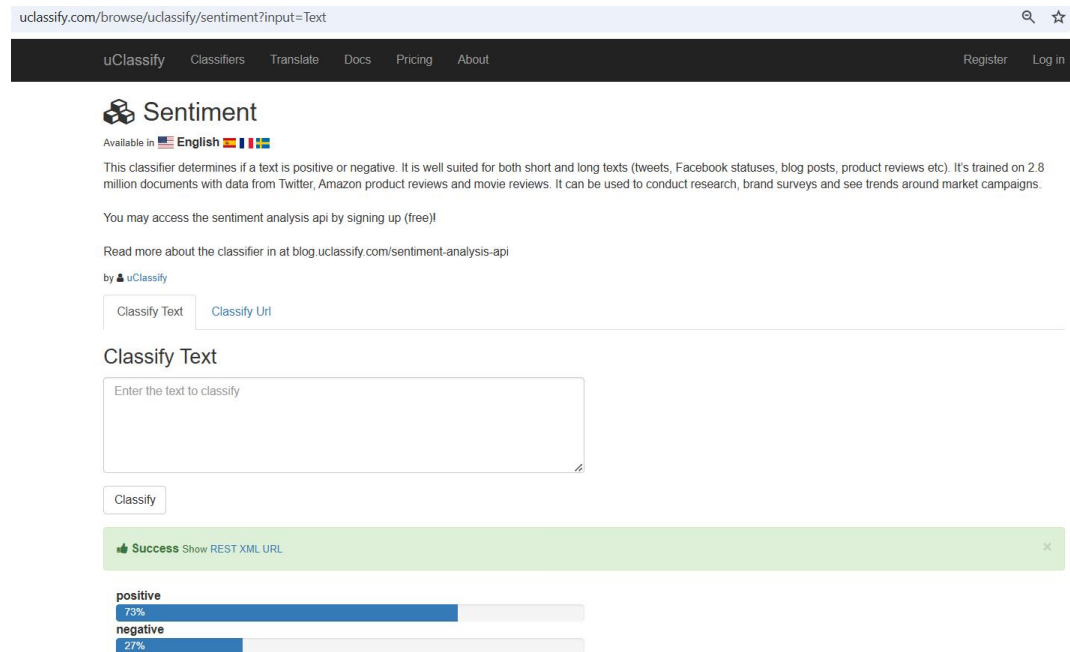
Ứng dụng của NLP

❖ Phân loại văn bản (text categorization)



Ứng dụng của NLP

❖ Phân loại văn bản (text categorization)

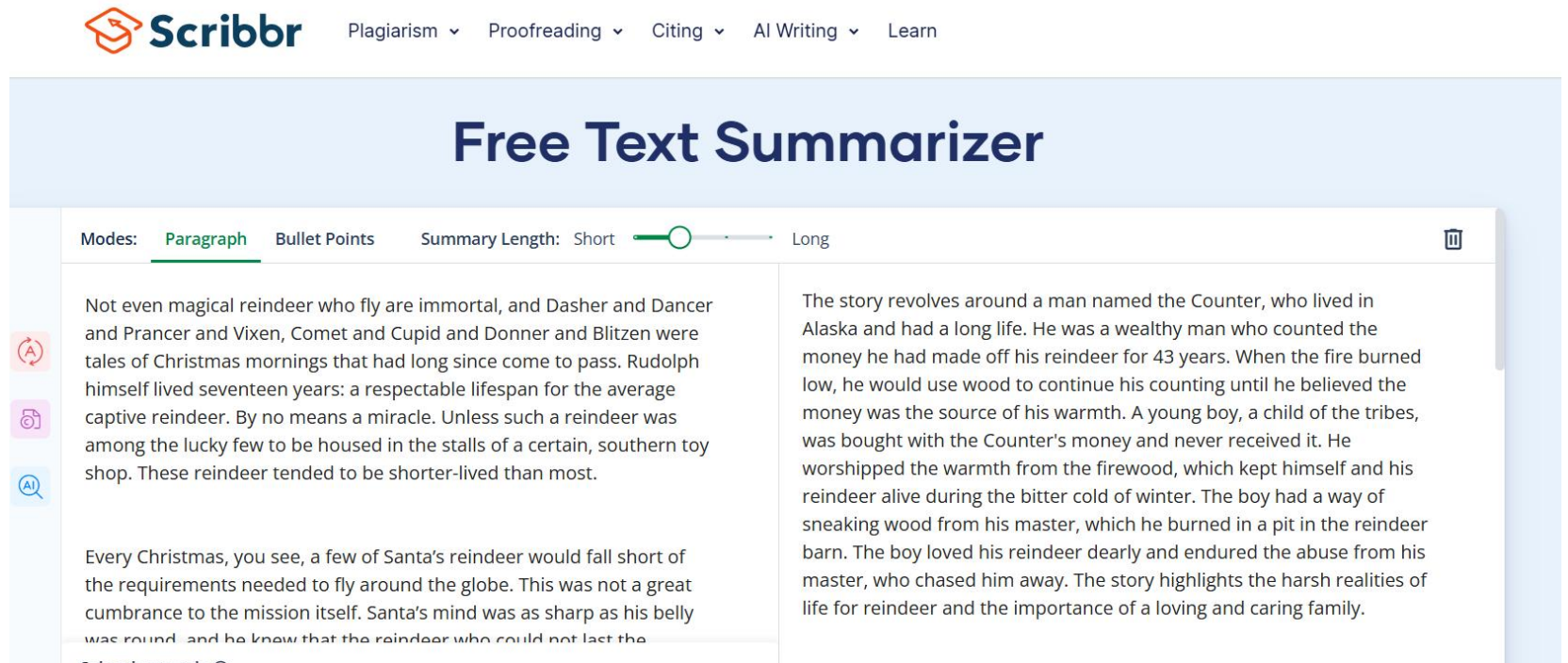


The screenshot displays the uClassify Sentiment classifier interface. The browser address bar shows the URL `uclassify.com/browse/uclassify/sentiment?input=Text`. The navigation bar includes links for uClassify, Classifiers, Translate, Docs, Pricing, About, Register, and Log in. The main heading is "Sentiment", with a note that it is available in English. A description states that the classifier determines if text is positive or negative, trained on 2.8 million documents from Twitter, Amazon, and movie reviews. Below this, there are tabs for "Classify Text" and "Classify Url". The "Classify Text" tab is active, showing a text input field with the placeholder "Enter the text to classify". A "Classify" button is positioned below the input field. A green success message bar indicates "Success" and provides a link to "Show REST XML URL". At the bottom, a horizontal bar chart displays the classification results: "positive" at 73% and "negative" at 27%.

Sentiment	Percentage
positive	73%
negative	27%

Ứng dụng của NLP


❖ Tóm tắt (summarization)



The screenshot displays the Scribbr website's 'Free Text Summarizer' tool. The interface includes a navigation bar with the Scribbr logo and links for Plagiarism, Proofreading, Citing, AI Writing, and Learn. The main heading is 'Free Text Summarizer'. Below this, there are controls for 'Modes' (Paragraph, Bullet Points) and 'Summary Length' (Short, Long), with a slider set to 'Short'. The tool is processing a text input on the left and displaying the summarized output on the right. The input text is a paragraph about reindeer, and the output is a shorter summary of the same text.

Scribbr Plagiarism ▾ Proofreading ▾ Citing ▾ AI Writing ▾ Learn

Free Text Summarizer

Modes: Paragraph Bullet Points Summary Length: Short Long 

Not even magical reindeer who fly are immortal, and Dasher and Dancer and Prancer and Vixen, Comet and Cupid and Donner and Blitzen were tales of Christmas mornings that had long since come to pass. Rudolph himself lived seventeen years: a respectable lifespan for the average captive reindeer. By no means a miracle. Unless such a reindeer was among the lucky few to be housed in the stalls of a certain, southern toy shop. These reindeer tended to be shorter-lived than most.

Every Christmas, you see, a few of Santa's reindeer would fall short of the requirements needed to fly around the globe. This was not a great cumbrance to the mission itself. Santa's mind was as sharp as his belly was round, and he knew that the reindeer who could not last the

The story revolves around a man named the Counter, who lived in Alaska and had a long life. He was a wealthy man who counted the money he had made off his reindeer for 43 years. When the fire burned low, he would use wood to continue his counting until he believed the money was the source of his warmth. A young boy, a child of the tribes, was bought with the Counter's money and never received it. He worshipped the warmth from the firewood, which kept himself and his reindeer alive during the bitter cold of winter. The boy had a way of sneaking wood from his master, which he burned in a pit in the reindeer barn. The boy loved his reindeer dearly and endured the abuse from his master, who chased him away. The story highlights the harsh realities of life for reindeer and the importance of a loving and caring family.

Ứng dụng của NLP

❖ Dịch máy (machine translation)

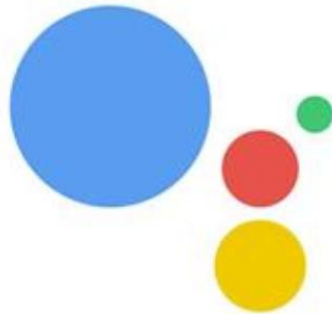
The screenshot displays the DeepL Translator web application. At the top, the navigation bar includes the DeepL logo, a 'Translator' dropdown menu, and links for 'Products', 'Solutions', 'Pricing', and 'Apps'. On the right side of the header, there are icons for help, a lightbulb, a 'Log in' button, and a 'Start free trial' button. Below the header, three main service buttons are visible: 'Translate text' (33 languages), 'Translate files' (.pdf, .docx, .pptx), and 'DeepL Write' (AI-powered edits). The main interface features a language selection bar with 'English (detected)' on the left and 'German' on the right, separated by a double-headed arrow. To the right of the language bar are 'Automatic' and 'Glossary' buttons. The central area is split into two panels. The left panel contains the text 'machine translation'. The right panel shows the translated text 'maschinelle Übersetzung'. Below this, under the heading 'Alternatives:', two other translations are listed: 'Maschinenübersetzung' and 'automatische Übersetzung'. At the bottom of the interface, there is a row of icons for voice input, text input, undo, redo, and a final action button. On the bottom right, there are icons for liking, commenting, editing, copying, and sharing.

Ứng dụng của NLP

❖ Trả lời câu hỏi



Siri



Google Assistant



Hey Cortana



Ứng dụng của NLP

❖ Phân tích cảm xúc (sentiment analysis)

text2data.com/Demo

☆ □ 📷 📄

Free sentiment analysis demo

Our demo service uses generic models trained on real user's comments, product, service opinions. In order to get specific results that are tailored to your domain, please consider training your own [sentiment model](#).

Please enter your text in **english*** for analysis or leave default one.

I purchased a larger one for bedroom and it arrived with a busted screen, so I ordered a replacement and got it on Friday. Took it out and set it up. NO picture - only static with a BLACK screen. It was hooked to Direct TV so we knew there was a problem when there was no picture and only static. It wouldn't respond to remote buttons or the buttons on the TV itself - definitely a problem. I called LG customer service and we performed a couple of their tests recommendations

☒ Twitter-like content ⓘ

🔗 SHARE THIS ANALYSIS

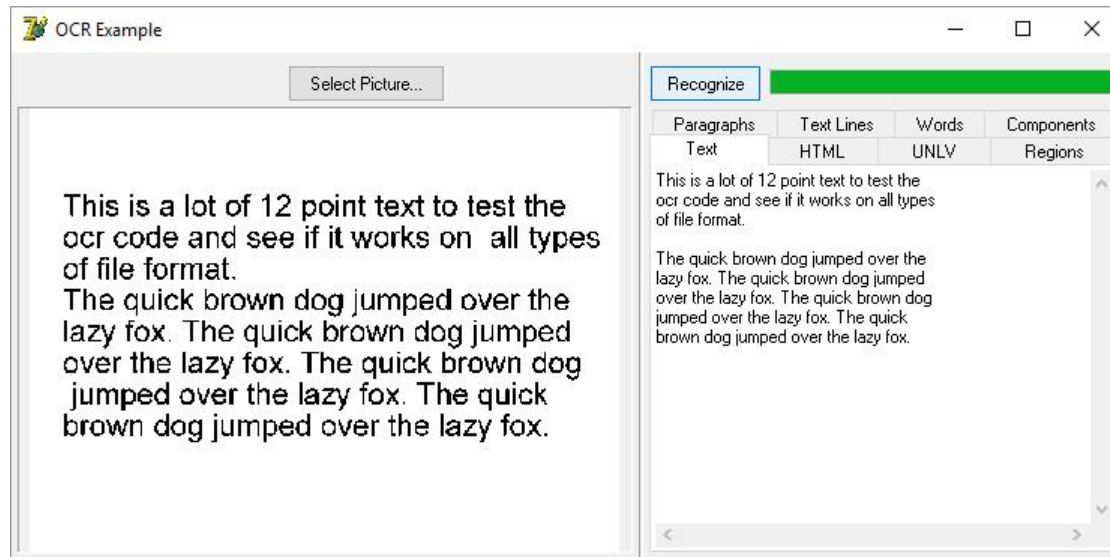
RUN ANALYSIS

I purchased a larger one for bedroom and it arrived with a busted screen, so I ordered a replacement and got it on Friday. Took it out and set it up. NO picture - only static with a BLACK screen. It was hooked to direct TV so we knew there was a problem when there was no picture and only static. It wouldn't respond to remote buttons or the buttons on the TV itself - definitely a problem. I called LG customer service and we performed a couple of their tests recommendations and finally got voice sound but still no picture, then we lost the voice again. The Customer Service lady told me this LG was defective. VERY disappointing to say the least - to receive not one but 2 broke/defective TV's. I'm ready to get my money back and try another brand. And I really do like my smaller LG so this is even more upsetting!

there was no picture still no
picture lost the voice static
least defective
disappointing

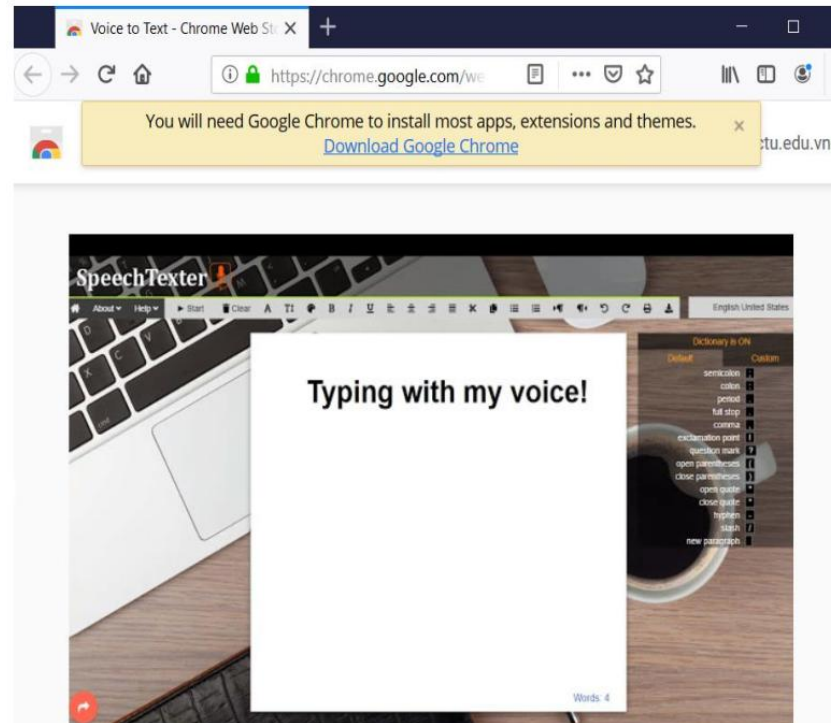
Ứng dụng của NLP


❖ Nhận dạng ký tự quang học (optical character recognition)



Ứng dụng của NLP

❖ Nhận dạng giọng nói (speech recognition)

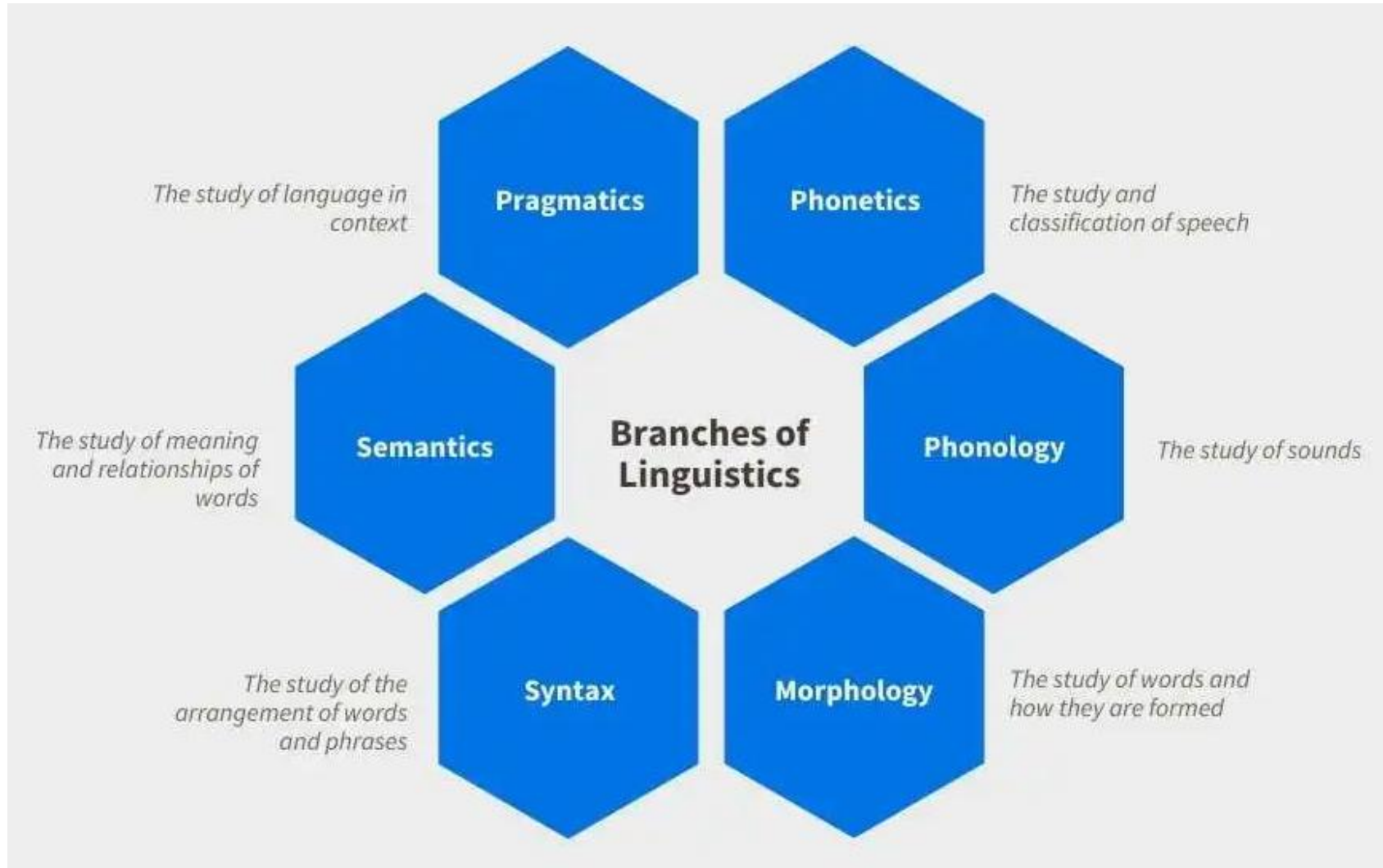




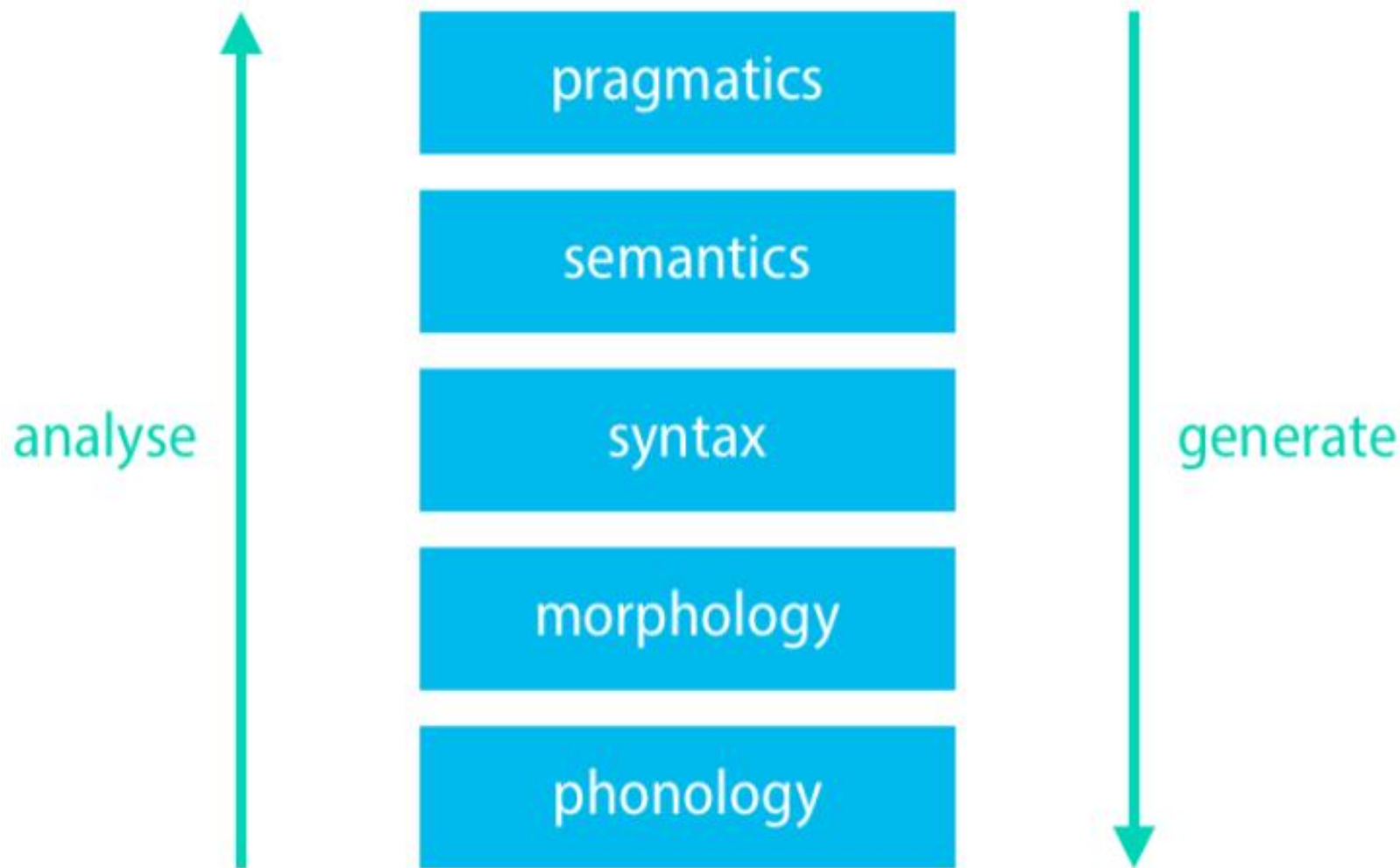
Kiến thức cơ bản về ngôn ngữ học

Xử lý ngôn ngữ tự nhiên

Ngôn ngữ học (Linguistics)




Các mức độ phân tích ngôn ngữ



Các mức độ phân tích ngôn ngữ



- ❖ Hệ thống âm vị (phonology)
 - ❖ Hình thái học (morphology)
 - ❖ Cú pháp (syntax)
 - ❖ Ngữ nghĩa (semantics)
 - ❖ Ngữ dụng (pragmatics)
- 

Âm vị học (Phonology)

- ❖ Âm vị học (phonology): nghiên cứu các quy tắc tổ chức các mẫu âm thanh trong ngôn ngữ loài người
 - Tiếng Việt có 16 âm vị nguyên âm và 23 âm vị phụ âm
 - Âm vị phụ âm: /f/, /t'/, /c/, /ɲ/,....
 - Âm vị nguyên âm: /i/, /e/, /ɛ/...
- ❖ Âm vị học (phonology) khác với ngữ âm học (phonetics) — nghiên cứu về việc tạo ra, truyền tải và nhận thức âm thanh mà không cần kiến thức trước về ngôn ngữ đang nói.

Hình thái học (Morphology)

❖ Hình thái học (morphology)

- nghiên cứu cách các từ được cấu tạo bởi các hình vị (morpheme), đơn vị nhỏ nhất có nghĩa trong ngôn ngữ
- ❖ Cấu trúc của một từ bao gồm một số hình vị
 - một gốc (root) hoặc thân từ (stem)
 - không hoặc nhiều tiếp vị (affix) như tiền tố (prefix) và hậu tố (suffix)
 - Ví dụ: draw, draw + s, draw + ing, un+ draw + able

Cú pháp (Syntax)

❖ Cú pháp (syntax)

- Nghiên cứu các quy tắc và ràng buộc chi phối cách các từ được sắp xếp thành câu

❖ Từ loại (part of speech):

- Các nhóm từ vai trò tương tự trong cấu trúc cú pháp của câu
- Có thể được định nghĩa:
 - ✓ Theo phân bố (distributionally): Kelly saw {errors, bugs, mountain} before we did.
 - ✓ Theo chức năng (functionally): động từ = vị ngữ, danh từ = bổ ngữ, trạng từ = bổ nghĩa cho động từ....

Cú pháp (Syntax)

❖ Từ loại (part of speech):

➤ Thẻ lớp mở (Open class tags)

- ✓ bao gồm danh từ (noun), động từ (verb), tính từ (adjective), trạng từ (adverb)
- ✓ các từ mới thường được thêm vào các lớp này
- ✓ đôi khi được gọi ***từ nội dung (content words)***

➤ Thẻ lớp đóng (Close class tags)

- ✓ bao gồm các từ hạn định (determiner), giới từ (preposition), liên từ (conjunction)...
- ✓ các lớp từ ít khi nhận thêm từ mới
- ✓ đôi khi được gọi ***từ chức năng (function words)***

Cú pháp (Syntax)

❖ Từ loại (part of speech):

➤ Cấu trúc cú pháp (Syntactic structure):

một số các biểu diễn cấu trúc cú pháp phổ biến

✓ cấu trúc cụm từ (phrase structure): biểu diễn tương tự cây với

❑ nút lá (leaf node) biểu diễn các từ trong câu

❑ nút trong (internal node) biểu diễn các cụm từ (phrase)

✓ cây phụ thuộc (dependency tree)

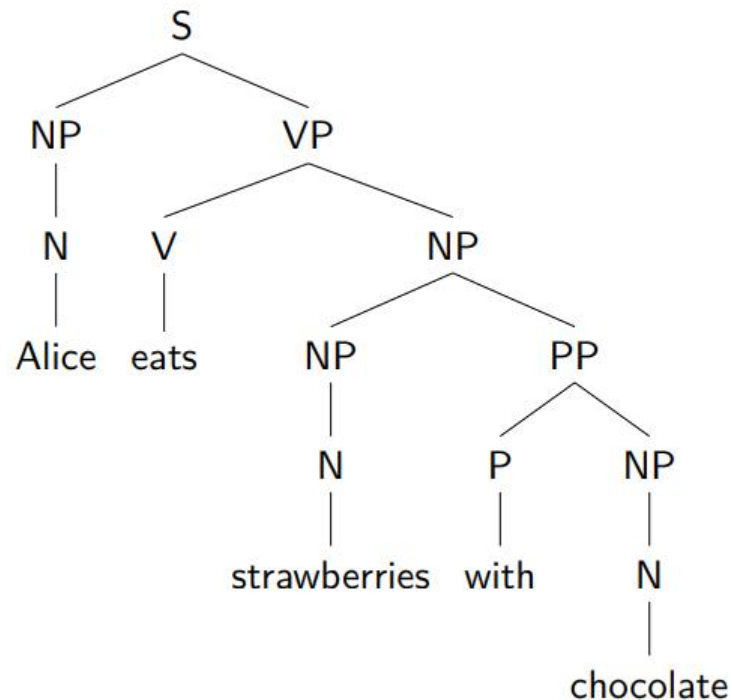
Cấu trúc cụm từ (phrase structure)

❖ Cấu trúc cụm từ sử dụng một số nhãn như

PoS tags	Phrase tags
—	S = Sentence
N = Noun	NP = Noun Phrase
V = Verb	VP = Verb Phrase
P = Preposition	PP = Prepositional Phrase
A = Adjective	AP = Adjectival Phrase
Det = Determiner	—
⋮	⋮

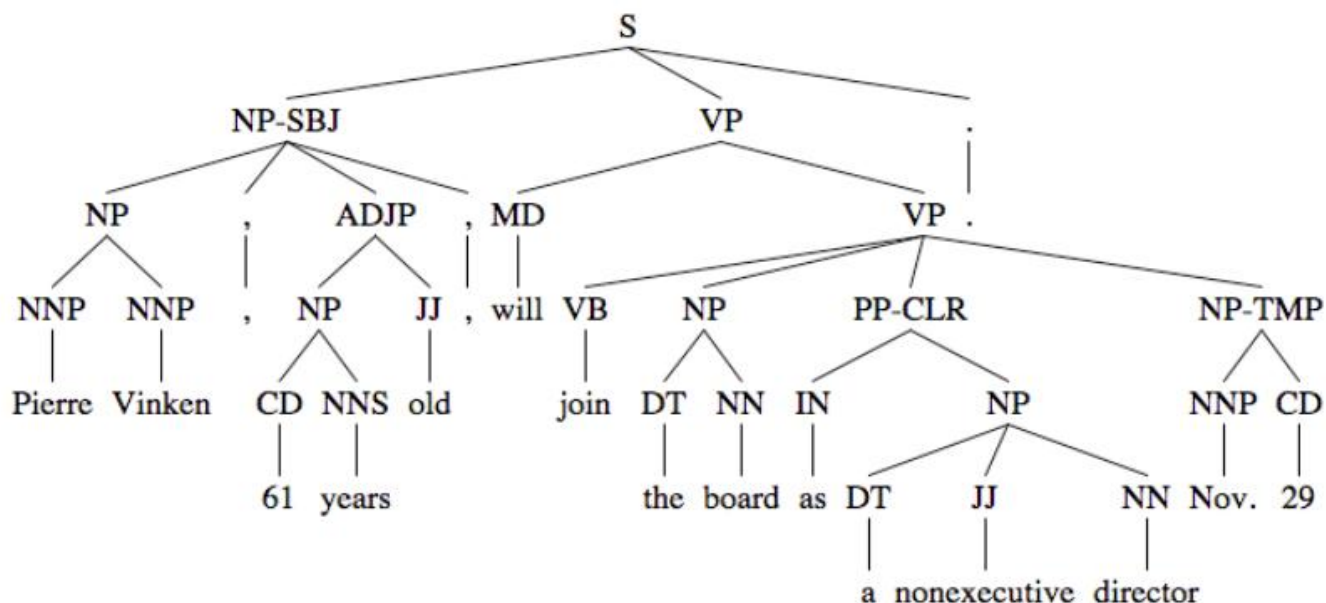
Cấu trúc cụm từ (phrase structure)

❖ Ví dụ: Alice eats strawberries with chocolate.



Cấu trúc cụm từ (phrase structure)

❖ **Ví dụ:** Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.



Cú pháp (Syntax)

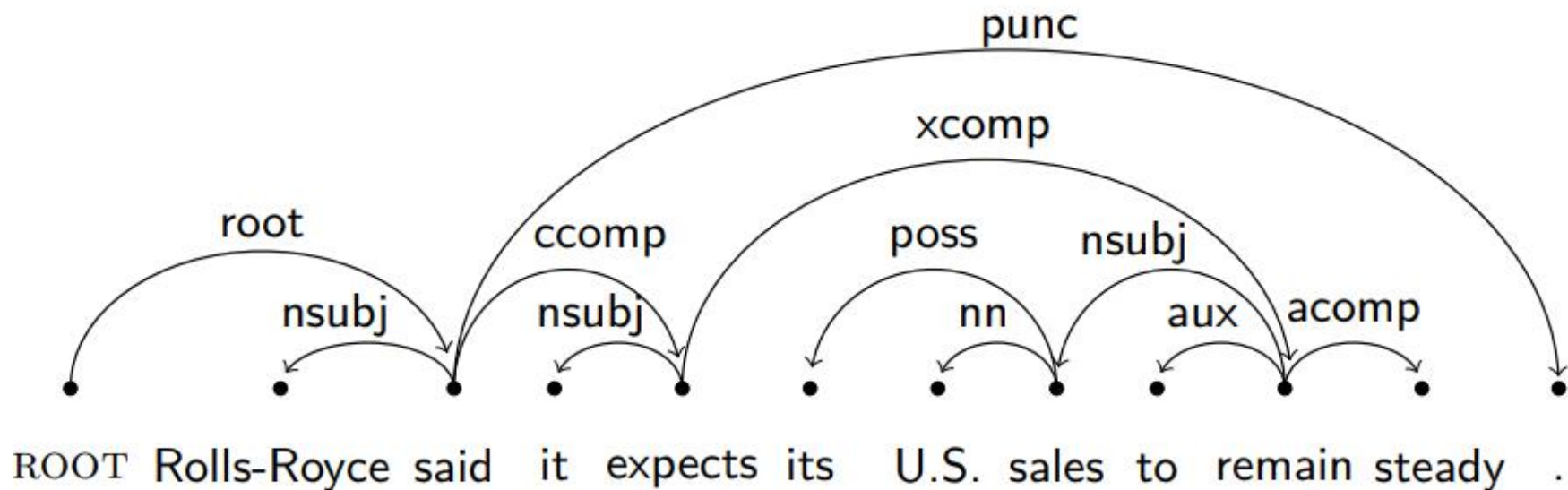
❖ Từ loại (part of speech):

➤ Cấu trúc cú pháp (Syntactic structure):

- ✓ cấu trúc cụm từ (phrase structure)
- ✓ cây phụ thuộc (dependency tree): biểu diễn tương tự cấu trúc cây
 - nút biểu diễn từ (và dấu câu) trong câu
 - câu (arc) biểu diễn mối quan hệ ngữ pháp giữa một từ đầu (head) và một từ phụ thuộc (dependent)

Cây phụ thuộc (dependency tree)

❖ Ví dụ: Rolls–Royce said it expects its U.S. sales to remain steady.



Ngữ nghĩa (semantics)

❖ Ngữ nghĩa học (semantics):

- nghiên cứu về ý nghĩa của các biểu thức ngôn ngữ như từ, cụm từ và câu
- trọng tâm là ý nghĩa thông thường/trừu tượng của các biểu thức chứ không phải ý nghĩa trong một ngữ cảnh cụ thể

❖ Ngữ nghĩa từ vựng (lexical semantics):

- nghiên cứu về nghĩa của từ
- **Cấu trúc ngữ nghĩa bên trong (internal semantic structure)** của một từ: sự tương đồng với các từ khác
 - ✓ Thay thế từ này bằng các từ khác có nghĩa như thế nào?
- **Cấu trúc ngữ nghĩa bên ngoài (external semantic structure)** của một từ: khả năng kết hợp với từ khác
 - ✓ Kết hợp từ này với các từ khác có nghĩa như thế nào?

Ngữ dụng (pragmatics)

❖ Ngữ dụng (pragmatics)

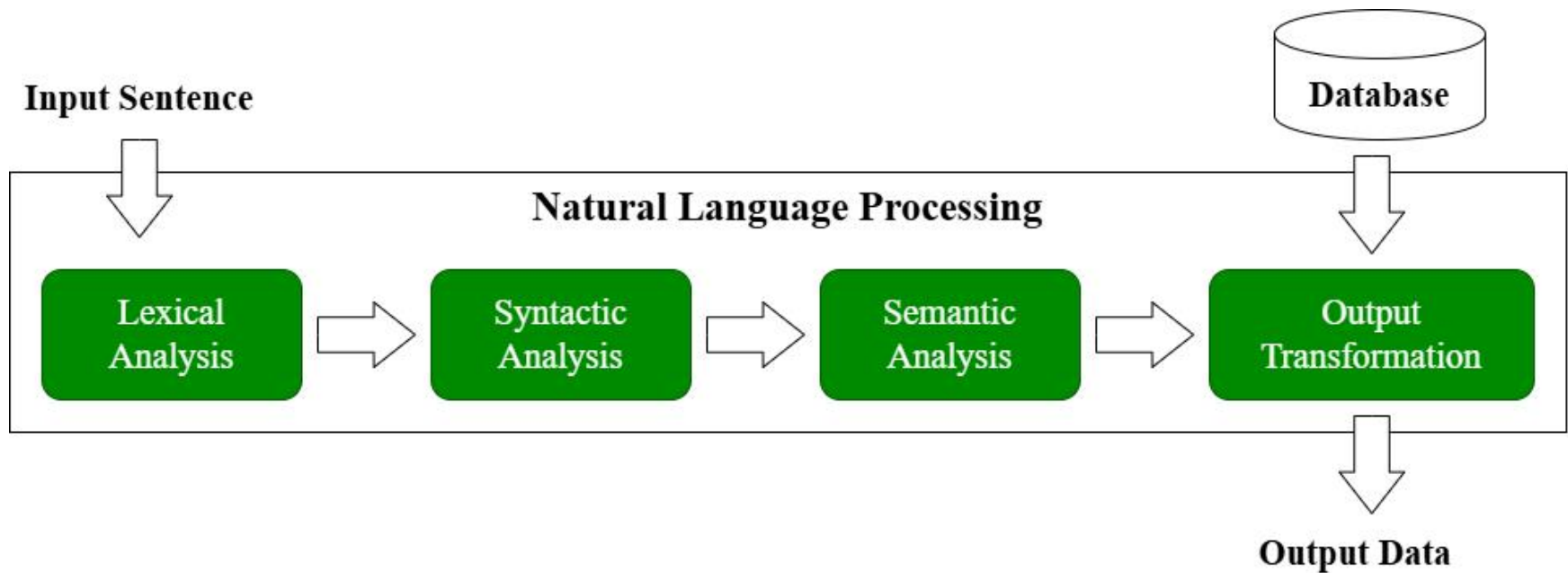
- nghiên cứu cách các biểu thức ngôn ngữ với ngữ nghĩa của chúng được sử dụng cho mục đích giao tiếp cụ thể
- nghiên cứu cách thức *ngữ cảnh (context)* ảnh hưởng đến ngữ nghĩa trong từng trường hợp cụ thể
- một khái niệm quan trọng trong ngữ dụng học là *hành động lời nói (speech act)*, mô tả một hành động thực hiện thông qua lời nói

The background of the slide is a dark, textured surface covered with various characters, numbers, and symbols in a glowing, ethereal blue and orange light. The characters appear to be floating or falling, creating a sense of dynamic movement. The overall aesthetic is high-tech and digital.

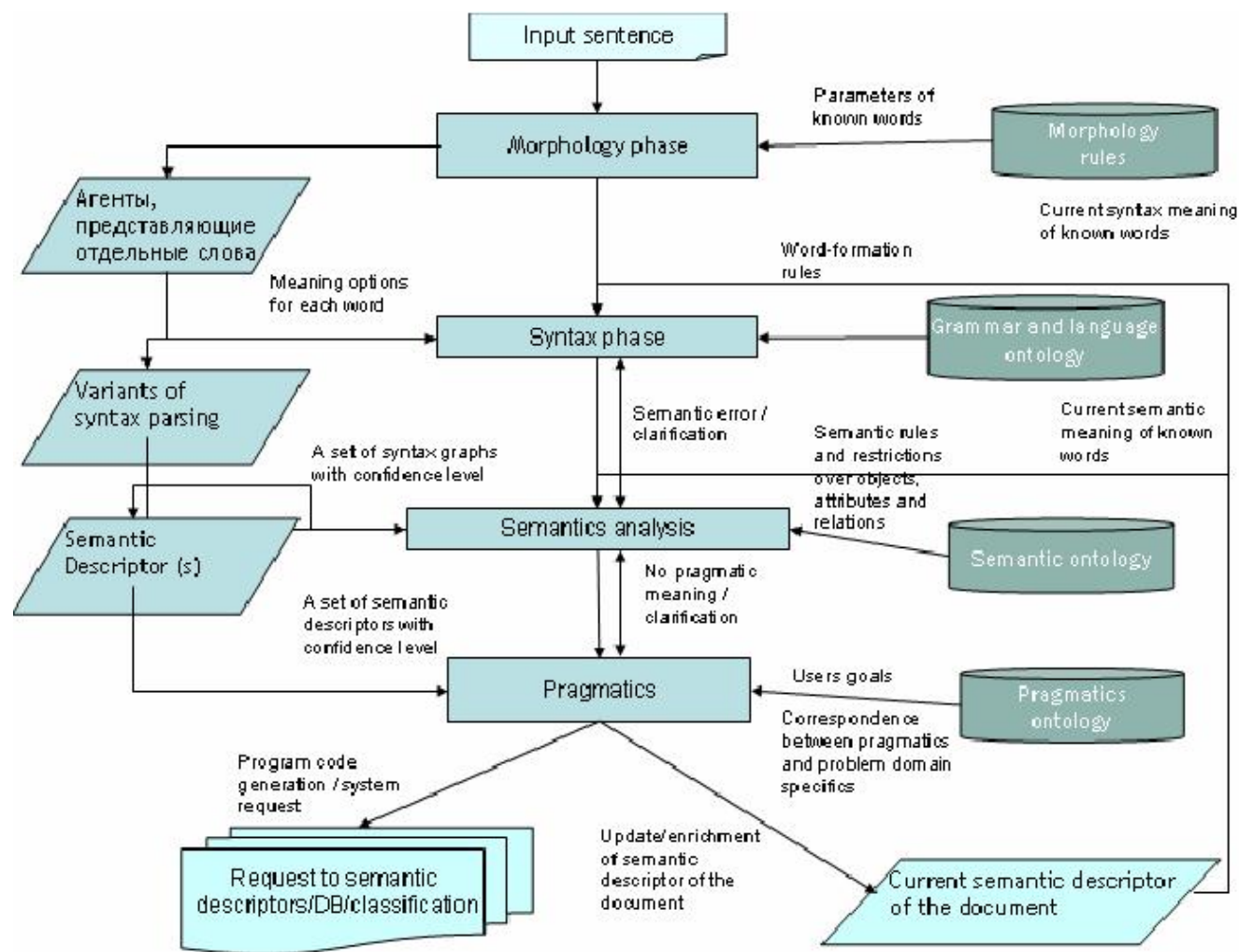
KỸ THUẬT XỬ LÝ NGÔN NGỮ TỰ NHIÊN

XỬ LÝ NGÔN NGỮ TỰ NHIÊN

Tác vụ NLP (NLP Tasks)



Sơ đồ tổng quát của các thuật toán NLP



Chia câu (Sentence splitting)

❖ Chia (split) văn bản thành các câu (sentence)

We are talking about pens.
He said "This is a pen. I like it".
I could relate to that statement.

Result is:

Paragraph

Sentences

We are talking about pens.

He said "This is a pen. I like it".

context

I could relate to that statement.

Gán nhãn từ loại (Part-of-speech tagging)

❖ Gán nhãn cú pháp cho mỗi từ trong câu

Parts-of-speech.Info

POS tagging

[about Parts-of-speech.Info](#)

Enter a complete sentence (no single words!) and click at "POS-tag!". The tagging works better when grammar and orthography are correct.

Text:

We are talking about pens . He said " This is a pen . I like it " . I could relate to that statement .

 Edit text



English



Adjective

Adverb

Conjunction

Determiner

Noun

Number

Preposition

Pronoun

Verb

Parsing

❖ Xây dựng cây cú pháp của một câu

The Georgetown experiment in 1954 involved fully automatic translation of more than sixty Russian sentences into English.

Person

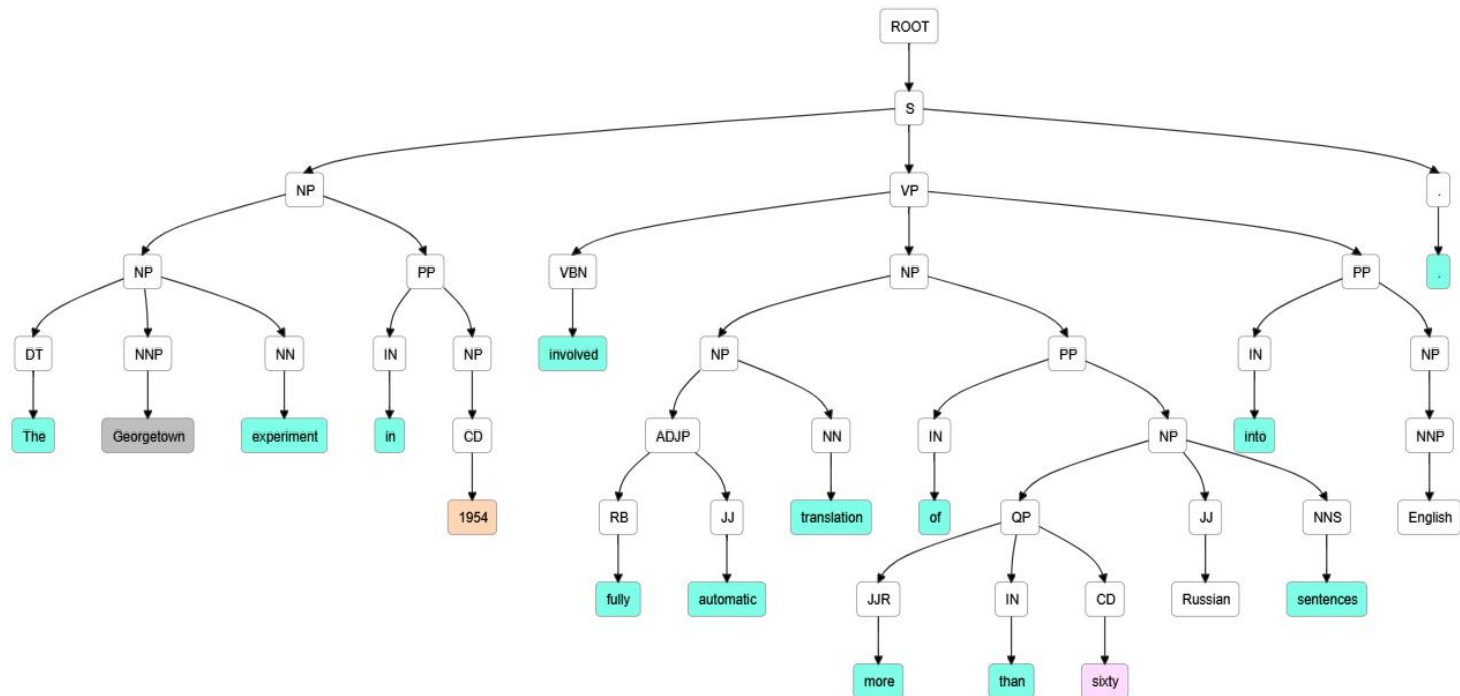
Date

Organization

Location

Ordinal

Number



Nhận dạng thực thể (Named-entity recognition)

- ❖ Xác định các loại thực thể được xác định trước trong một câu

!/?# Named Entity Recognition

Input a piece of text:

He and Warren had traveled together for four days .

52/800

Recognize

Result

PRP CC PERSON
NNP VBD VBN RB IN CD NNS .
He and Warren had traveled together for four days .

Giải thích nghĩa của từ (Word sense disambiguation)

❖ Tìm ra nghĩa chính xác của một từ hoặc một thực thể

WordNet Search - 3.1

[- WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- [S:](#) (n) **wordnet** (any of the machine-readable lexical databases modeled after the Princeton WordNet)
- [S:](#) (n) **WordNet**, [Princeton WordNet](#) (a machine-readable lexical database organized by meanings; developed at Princeton University)

Gán nhãn vai trò ngữ nghĩa (semantic role labeling)

- ❖ Trích xuất đối tượng chủ ngữ – vị ngữ từ 1 câu
- ❖ Gán nhãn vai trò ngữ nghĩa còn được gọi là phân tích ngữ nghĩa nông

He and Warren had traveled together for four days.

50/800

Label

Result

Token	SRL PA1
He	← →ARG0
and	
Warren	
had	
traveled	→PRED
together	→ARGM-MNR
for	← →ARGM-TMP
four	
days	
.	

Những phương pháp tiếp cận chính

