



CSE: Faculty of Computer Science and Engineering  
Thuyloi University

---

# XỬ LÝ NGÔN NGỮ TỰ NHIÊN (NATURAL LANGUAGE PROCESSING)

*TS. Nguyễn Thị Kim Ngân*  
Email: ngannguyen@tlu.edu.vn

*Có tham khảo bài giảng của PGS.TS Nguyễn Thanh Hương, trường đại học Bách khoa Hà Nội*



# Xử lý ngôn ngữ tự nhiên

---

Xử lý ngôn ngữ tự nhiên (NLP) là một lĩnh vực con liên ngành của khoa học máy tính và trích xuất thông tin. Nó liên quan đến việc xử lý các tập dữ liệu ngôn ngữ tự nhiên, chẳng hạn như văn bản hoặc giọng nói bằng các phương pháp học máy. Mục tiêu là một máy tính có khả năng "hiểu" nội dung của tài liệu, bao gồm cả các sắc thái ngữ cảnh của ngôn ngữ bên trong chúng



# Thông tin môn học

---

- Là học phần chuyên ngành của sinh viên ngành Trí tuệ nhân tạo
- Môn học tiên quyết: Đại số tuyến tính, Giải tích, Xác suất thống kê, Học máy, Học sâu
- Kỹ năng lập trình: Python cơ bản



# Mục tiêu môn học

---

- Cung cấp các kiến thức cơ bản về
  - Xử lý ngôn ngữ tự nhiên
  - Một số bài toán cơ bản trong Xử lý ngôn ngữ tự nhiên
- Kỹ năng thực hành thuật toán học máy trên Python
  - Sinh viên cài đặt được một số bài toán Xử lý ngôn ngữ tự nhiên



# Đánh giá

---

- Điểm quá trình: 50%
  - Bài tập: 20%
  - Kiểm tra trên lớp: 20%
  - Vắng  $\leq 9$  tiết: 10%
- Thi cuối kỳ (vấn đáp): 50%



# Tài liệu tham khảo

---

- Jacob Eisenstein. Introduction to Natural Language Processing. The MIT Press. 2019  
<https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf>
- Một số tài liệu tham khảo khác:
  - Phương pháp tách từ trong văn bản <https://phamdinhhkhanh.github.io/2019/04/29/ModelWord2Vec.html>
  - Sử dụng thư viện Transformers trong NLP <https://huggingface.co/learn/nlp-course/vi/chapter1/1>
  - Natural Language Processing with Transformers, Revised Edition <https://github.com/nlp-with-transformers/notebooks>
  - <https://www.oreilly.com/library/view/natural-language-processing/9781098136789/>
  - <https://github.com/nlp-with-transformers/notebooks>
  - <https://github.com/yhilpisch/aiif>



# Ngôn ngữ lập trình python

---

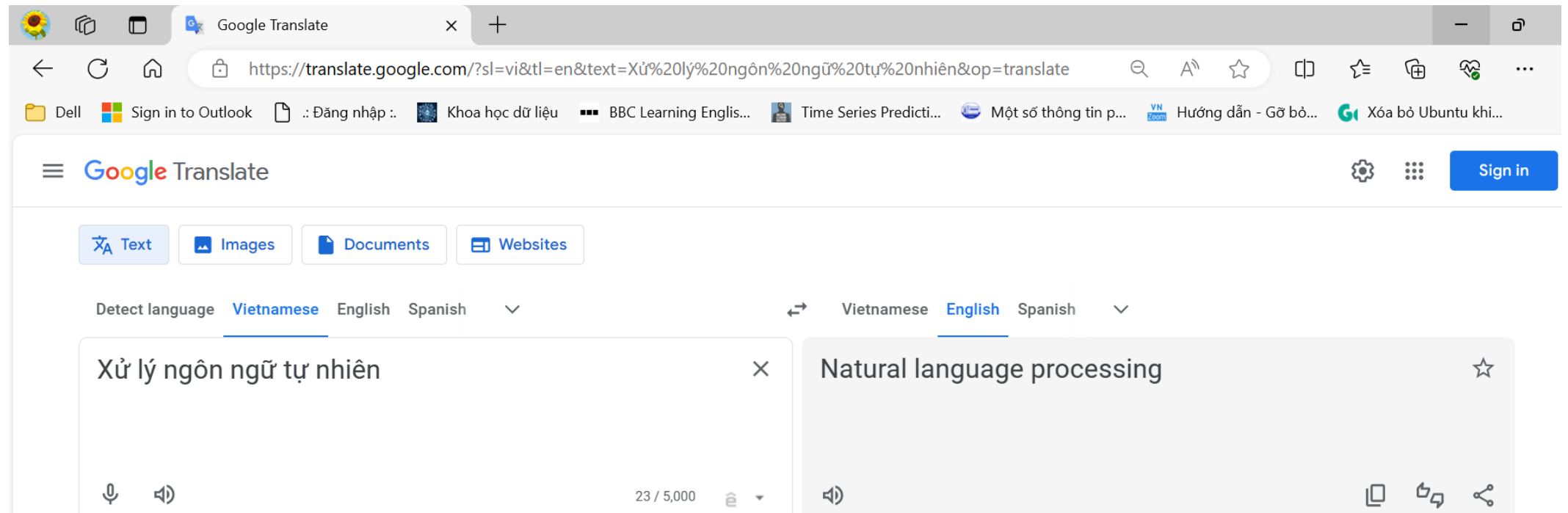
**Numpy** ([http:// www.numpy.org/](http://www.numpy.org/)): thư viện giúp xử lý các phép toán liên quan đến các mảng nhiều chiều, với các hàm gần gũi với đại số tuyến tính

**Tensorflow**( <https://www.tensorflow.org/> ): thư viện hỗ trợ các mô hình Deep Learning

**Transformers** (<https://huggingface.co/> ): thư viện hỗ trợ *pretrain* và *fine tuning the model*

# Một số ứng dụng của ngôn ngữ tự nhiên

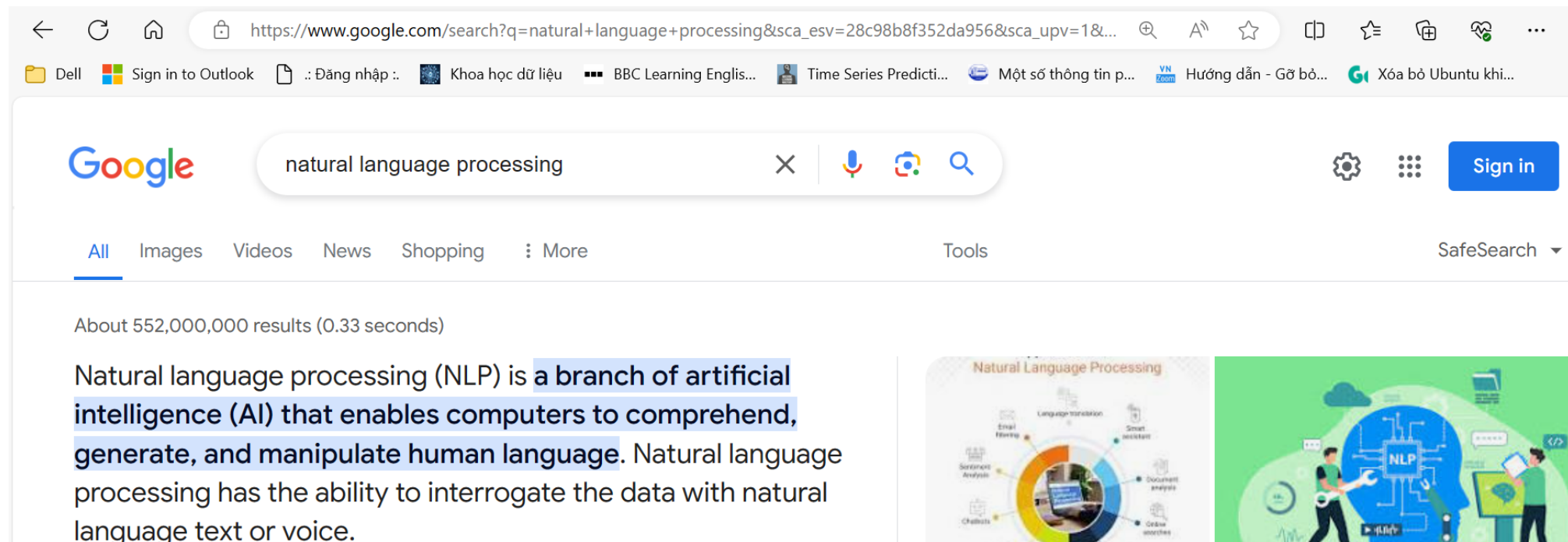
## Dịch máy





# Một số ứng dụng của ngôn ngữ tự nhiên

## Trích rút thông tin



The screenshot shows a Google search page with the query "natural language processing". The search results indicate approximately 552,000,000 results found in 0.33 seconds. The first result is a definition of NLP as a branch of artificial intelligence that enables computers to understand, generate, and manipulate human language. To the right of the text are two illustrative images: one showing a circular diagram of NLP applications like sentiment analysis and chatbots, and another showing a stylized head with a brain and NLP components.

← ↻ 🏠 [https://www.google.com/search?q=natural+language+processing&sca\\_esv=28c98b8f352da956&sca\\_upv=1&...](https://www.google.com/search?q=natural+language+processing&sca_esv=28c98b8f352da956&sca_upv=1&...) 🔍 🗨️ ☆ 📄 📌 📁 📧 ...

📁 Dell 📧 Sign in to Outlook 📄 : Đăng nhập : 📖 Khoa học dữ liệu 📄 BBC Learning Englis... 📄 Time Series Predicti... 📄 Một số thông tin p... 📄 VN Zoom 📄 Hướng dẫn - Gỡ bỏ... 📄 Xóa bỏ Ubuntu khi...



**Google** natural language processing ✕ 🗣️ 📷 🔍 ⚙️ ☰ Sign in

All Images Videos News Shopping ⋮ More Tools SafeSearch ▾

About 552,000,000 results (0.33 seconds)

Natural language processing (NLP) is a branch of artificial intelligence (AI) that enables computers to comprehend, generate, and manipulate human language. Natural language processing has the ability to interrogate the data with natural language text or voice.

**Natural Language Processing**



# Một số ứng dụng của ngôn ngữ tự nhiên

## Trích rút thông tin

October 14, 2002, 4:00 a.m. PT

For years, [Microsoft Corporation](#) CEO [Bill Gates](#) railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Today, Microsoft claims to "love" the open-source concept, by which software code is made public to encourage improvement and development by outside programmers. Gates himself says Microsoft will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said [Bill Veghte](#), a [Microsoft VP](#). "That's a super-important shift for us in terms of code access."

[Richard Stallman](#), [founder](#) of the [Free Software Foundation](#), countered saying...



NAME	TITLE	ORGANIZATION
<a href="#">Bill Gates</a>	CEO	<a href="#">Microsoft</a>
<a href="#">Bill Veghte</a>	VP	<a href="#">Microsoft</a>
<a href="#">Richard Stallman</a>	founder	<a href="#">Free Soft..</a>



# Một số ứng dụng của ngôn ngữ tự nhiên

---

Hỏi đáp

## Nói chuyện với Eliza

- E: how are you today?
- H: not so good
- E: why do you say not so good?
- H: i suffer from allergies?
- E: that's interesting
- H: i come out in spots
- E: i'm not sure i know what you're talking about
- H: spots on my face
- E: your face
- H: that's right
- E: is it really that right?

<https://en.wikipedia.org/wiki/ELIZA>



# Nghiên cứu Xử lý ngôn ngữ tự nhiên

---

- Nghiên cứu cách con người xác định từ
- Nghiên cứu cách con người phân tích câu
- Nghiên cứu cách con người học một ngôn ngữ
- Nghiên cứu cách ngôn ngữ tiến hóa



# Các chủ đề trong XLNNTN

---

- Mức phân tích: cú pháp, ngữ nghĩa, diễn ngôn, thực chứng, ...
- Các bài toán con: gán nhãn từ loại, PTCP, phân giải nhập nhằng từ, ...
- Thuật toán và phương pháp: dựa trên tập ngữ liệu, dựa trên tri thức, ...
- Các ứng dụng: trích rút thông tin, phản hồi thông tin, dịch máy, hỏi đáp, hiểu ngôn ngữ tự nhiên



# Các mức phân tích

---

- Morphology (hình thái học): cách từ được xây dựng, các tiền tố và hậu tố của từ
- Syntax (cú pháp): mối liên hệ về cấu trúc ngữ pháp giữa các từ và ngữ
- Semantics (ngữ nghĩa): nghĩa của từ, cụm từ, và cách diễn đạt
- Discourse (diễn ngôn): quan hệ giữa các ý hoặc các câu



# Hình thái học

---

Tiếng Anh: ngôn ngữ biến hình, đa âm tiết

- kick kicks kicked kicking kick, kicks, kicked, kicking
- sit, sits, sat, sitting
- murder, murders

Nhưng không luôn có quy tắc:

- gorge(v: nhồi nhét, n: những cái đã ăn, hẻm núi), gorgeous (rực rỡ)
- arm (cánh tay), army (quân đội)

Tiếng Việt: ngôn ngữ không biến hình, đơn âm tiết -> cần tách từ



# Tách từ

---

- Một câu có thể có n khả năng tách từ, nhưng chỉ 1 trong chúng là đúng
- Giải pháp đơn giản: lấy chuỗi âm tiết dài nhất bắt đầu từ vị trí hiện tại và có trong từ điển từ
- Vấn đề: chồng chéo từ
  - Học sinh | học sinh | học.
  - Học sinh | học | sinh học.

=> Liệt kê tất cả các khả năng có thể và thiết kế một giải pháp để lựa chọn cái tốt nhất





# Gán nhãn từ loại

---

The boy threw a ball to the brown dog.

- The/**DT** boy/**NN** threw/**VBD** a/**DT** ball/**NN** to/**IN** the/**DT** brown/**JJ** dog/**NN**./.

DT – determiner từ chỉ định

NN – noun, danh từ, số ít hoặc số nhiều

VBD – verb, past tense động từ, quá

IN – preposition giới từ

JJ – adjective tính từ

. – dấu chấm câu



# Gán nhãn từ loại

---

Con ngựa đá con ngựa đá.

Con ngựa/DT đá/ĐgT con ngựa/DT đá/DT.

Ông già đi nhanh quá.

Ông/ĐaT già/TT đi/Phó\_từ nhanh/TT quá/trạng\_từ.

Ông già/DT đi/ĐgT nhanh/TT quá/trạng\_từ.



# Ngữ pháp: nhập nhằng cấu trúc (từ loại)

---

Time flies like an arrow.

Time // flies (VBZ) like (IN, giới từ so sánh) an arrow.

Time flies (NNS)// like (VBP) an arrow.



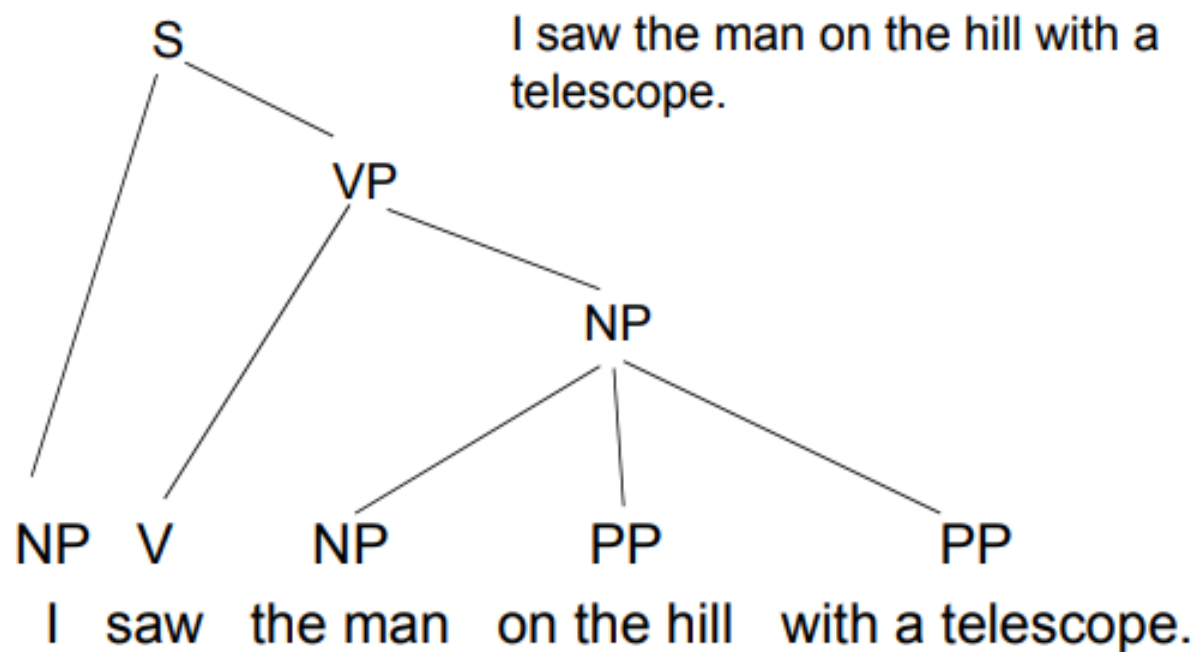
# Ngữ pháp: nhập nhằng cấu trúc (từ loại)

---

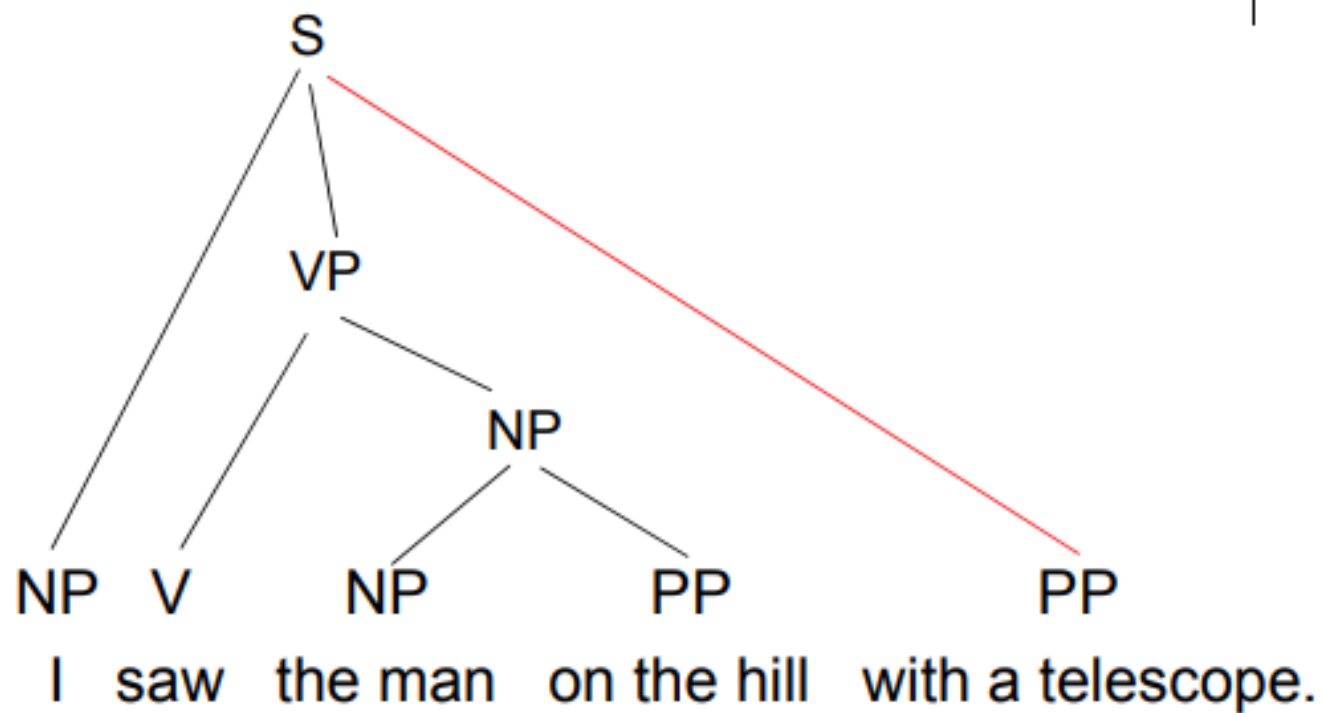
Ông già // đi nhanh quá.

Ông // già đi nhanh quá.

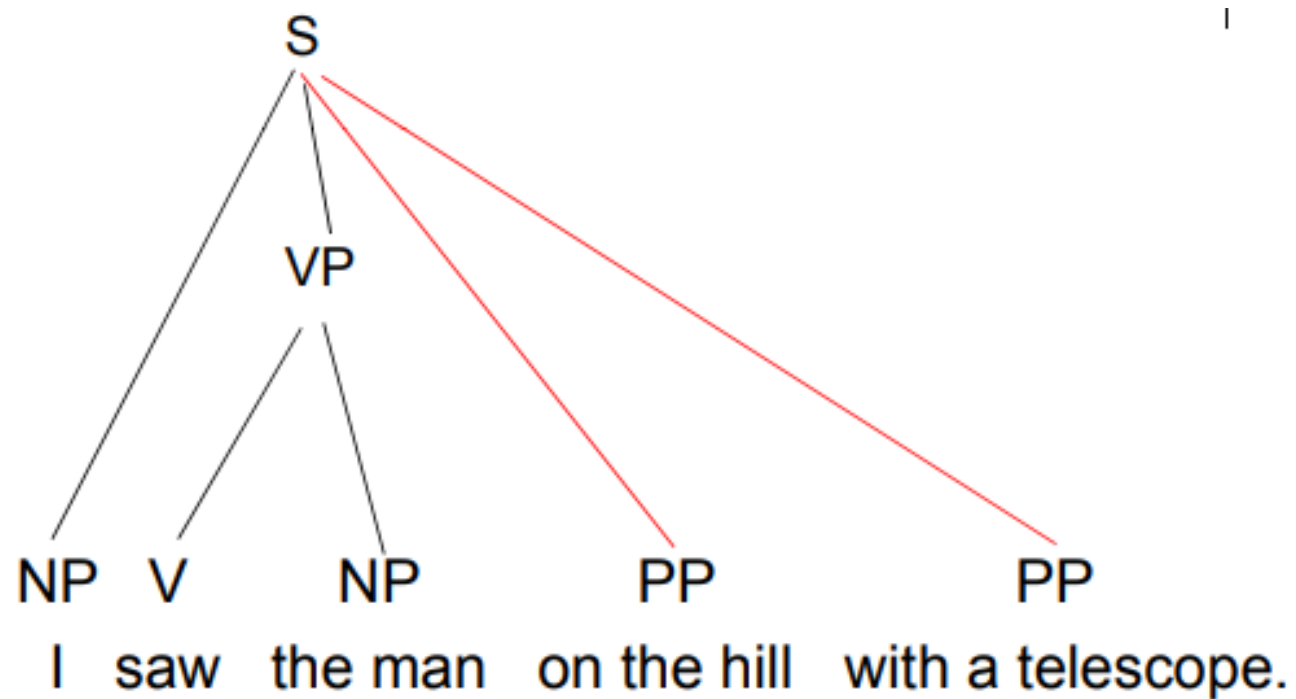
# Ngữ pháp: nhập nhằng cấu trúc (liên kết)



# Ngữ pháp: nhập nhằng cấu trúc (liên kết)



# Ngữ pháp: nhập nhằng cấu trúc (liên kết)





# Ngữ nghĩa: nhập nhằng mức từ vựng

---

I walked to the **bank** ...

of the river  
to get money.

I work for **John Hancock** ...

and he is a good boss.  
which is a good compan





# Diễn ngôn: đồng tham chiếu

---

President John F. Kennedy was assassinated. The president was shot yesterday.

Relatives said that John was a good father.

JFK was the youngest president in history.

His family will bury him tomorrow.

Friends of the Massachusetts native will hold a candlelight service in Mr. Kennedy's home town



# Vì sao XLNNTN khó?

---

- Ngôn ngữ tự nhiên:
  - Nhập nhằng tại mọi mức
  - Phức tạp và mờ
  - Liên quan lập luận về thế giới
- Không có tương ứng 1-1 với bất kỳ cách biểu diễn nào
- Ta cần biết cấu trúc dữ liệu và thuật toán để thực hiện, mặc dù có thể xảy ra bùng nổ tổ hợp ở bất cứ công đoạn xử lý nào



# Giải pháp

---

- Ta cần các công cụ nào?
  - Tri thức về ngôn ngữ
  - Tri thức về thể giới
  - Cách kết hợp các tri thức
- Giải pháp tiềm năng:
  - Các mô hình xác suất xây dựng từ dữ liệu
    - $P(\text{"maison"} \rightarrow \text{"house"})$  cao
    - $P(\text{"L'avocat general"} \rightarrow \text{"the general avocado"})$  thấp



# Các bài toán cơ bản

---

- Tách từ
- Gán nhãn từ loại
- Phân tích ngữ pháp
- Hiểu ngữ nghĩa