

**Section 0:**  
**Advanced Natural Languages Processing**



Lecturer: Assoc.Prof. Dr. Dinh Dien

[ddien@fit.hcmus.edu.vn](mailto:ddien@fit.hcmus.edu.vn)

# Lecturer

Đinh Diên 丁田

Dinh Dien  
Динх Диэн

딘 디엔

ディンディエン



Teaching Assistant: Dr. Buu Long – Dr. An Vinh

# **Natural Language Processing (NLP)**

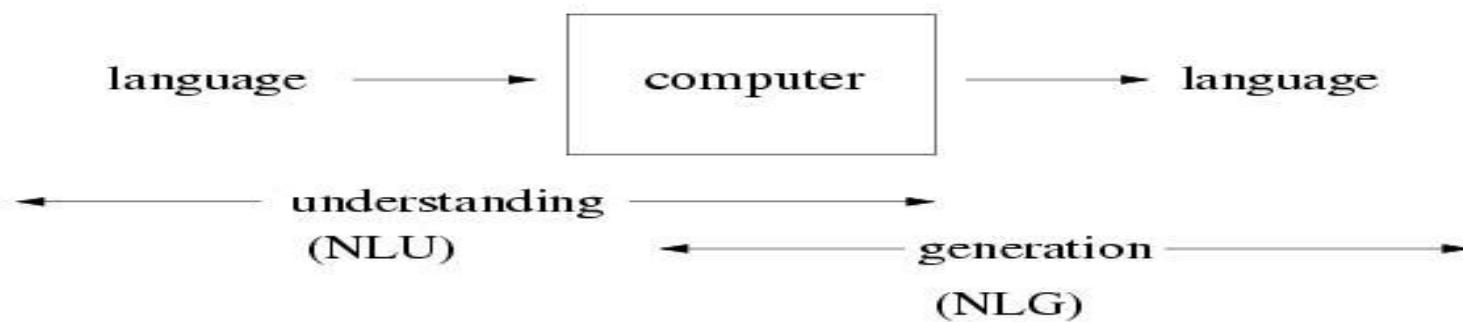
- 1. What is NLP ?**
- 2. Why NLP is hard?**
- 3. NLP applications**
- 4. What will this course be about?**
- 5. Textbooks - Website - dataset**
- 6. Grading:**
  - Mid-term: Seminar (40%)
  - Project (topics): oral examination (60%)
  - Bonus: unlimited

# What is NLP?

Using computer (Artificial Intelligence) to deal with human languages.

## What is Natural Language Processing?

computers using natural language as input and/or output



# Why NLP is hard?

- Reason (1) – human language is **ambiguous**:
- Ex1 (pronoun resolution):
  - Jack drank the wine on the table. *It* was red and round.
  - Jack saw Sam at the party. *He* went back to the bar to get another drink.
  - Jack saw Sam at the party. *He* clearly had drunk too much.
- Ex2: PrePosition Attachment:
  - I ate **the bread with** pecans.
  - I **ate the bread with** fingers.

# Why NLP is hard?

- Reason (2) – requires reasoning beyond what is explicitly mentioned (*A,B*) , and some of the reasoning requires world knowledge (*C*).
- *Ex: I couldn't submit my homework because my horse ate it.*

Implies that...

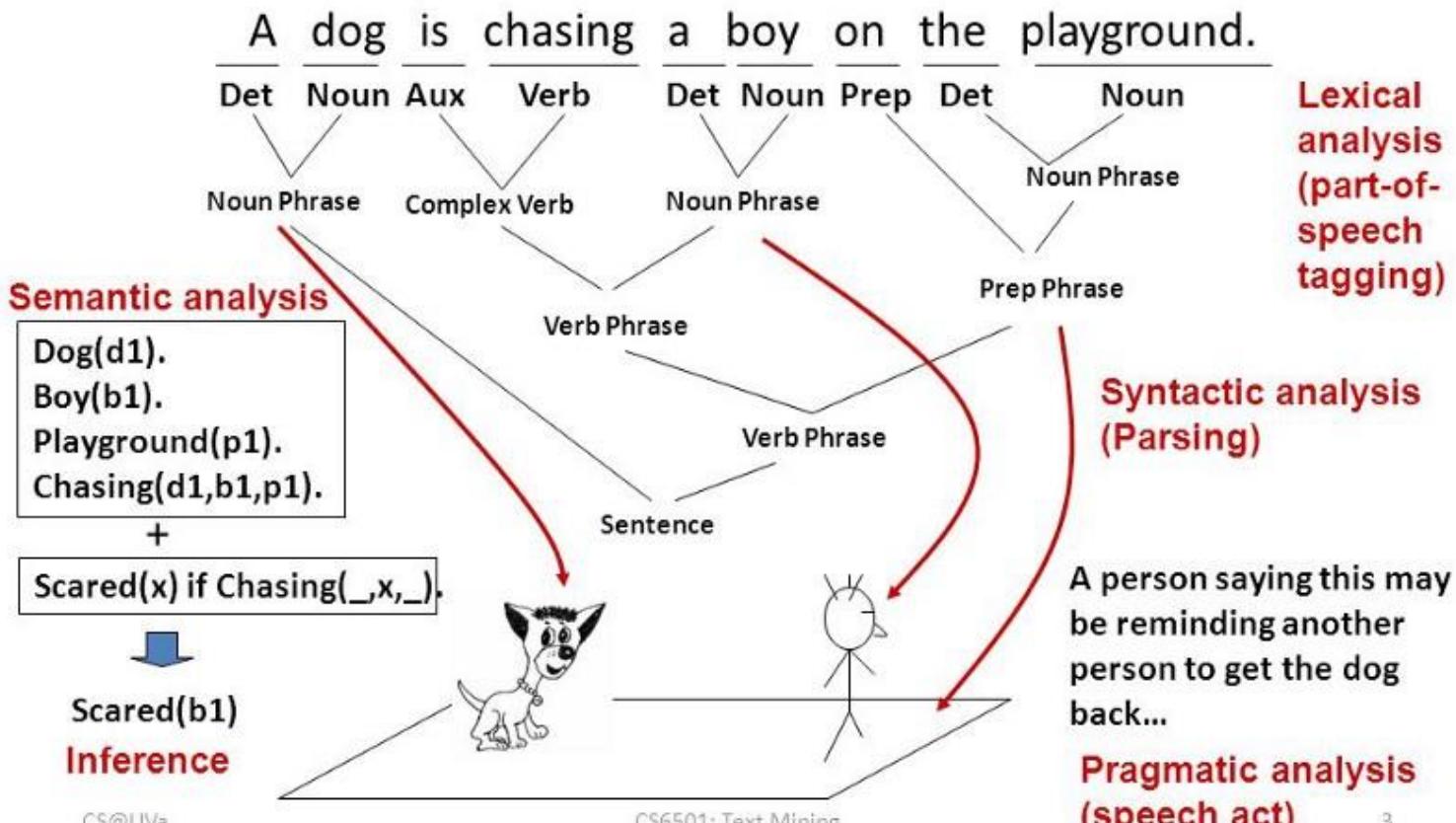
- *A. I have a horse.*
- *B. I did my homework.*
- *C. My homework was done on a soft object (such as papers) as opposed to a hard/heavy object (such as a computer). – it's more likely that my horse ate papers than a computer.*

# NLP APPLICATIONS

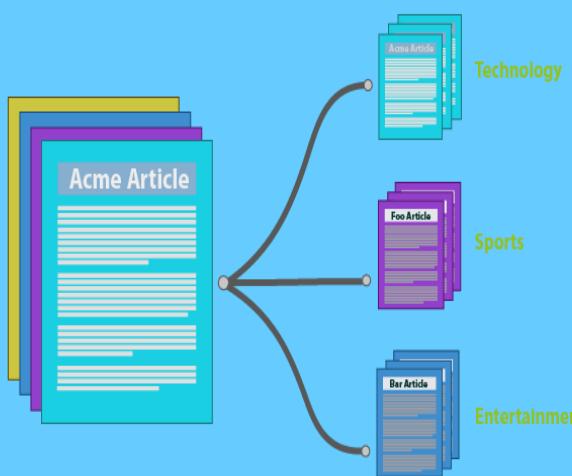
## 1. Linguistics

analysis:

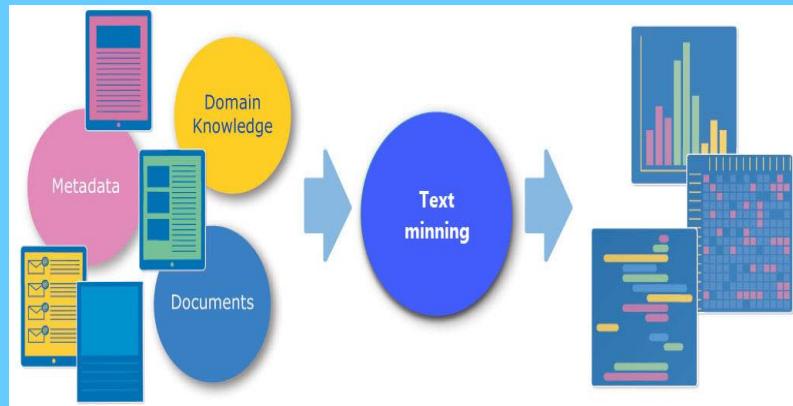
## An example of NLP



## 2. Text classification:

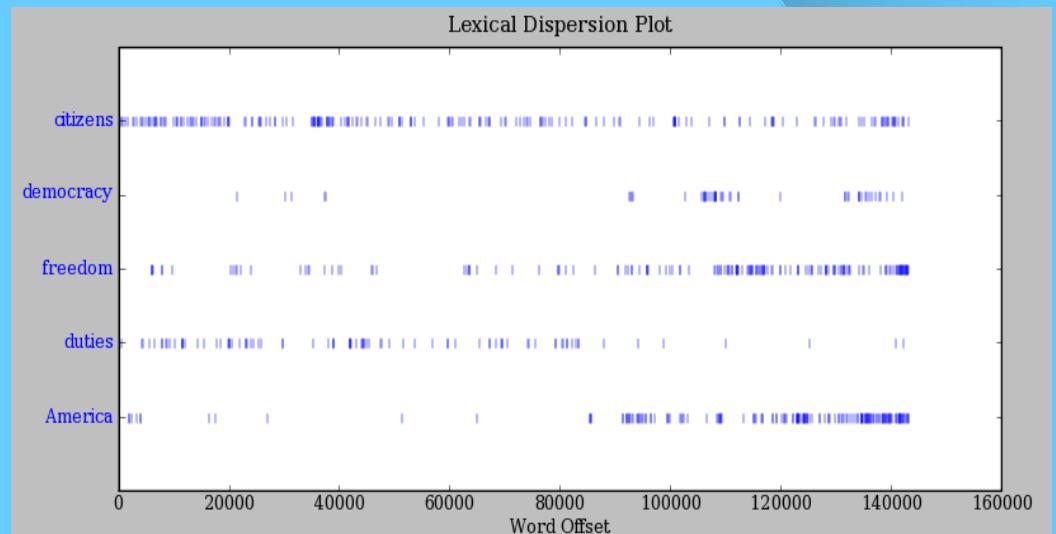
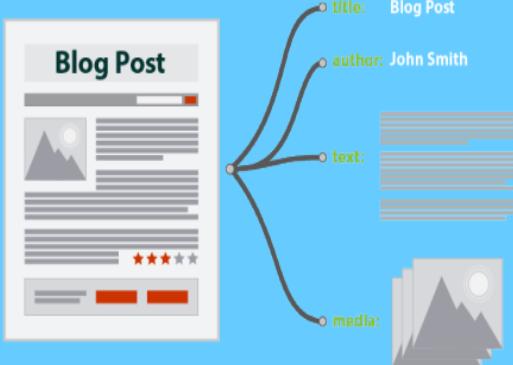


## Web Mining



After 9-11: “We”> “I”

## 3. Text mining:



# Information Extraction & Sentiment Analysis



Attributes:  
zoom  
affordability  
size and weight  
flash  
ease of use

## Size and weight

- ✓ nice and compact to carry!
- ✓ since the camera is small and light, I won't need to carry around those heavy, bulky professional cameras either!
- ✗ the camera feels flimsy, is plastic and very light in weight
- ✗ you have to be very delicate in the handling of this camera



DU

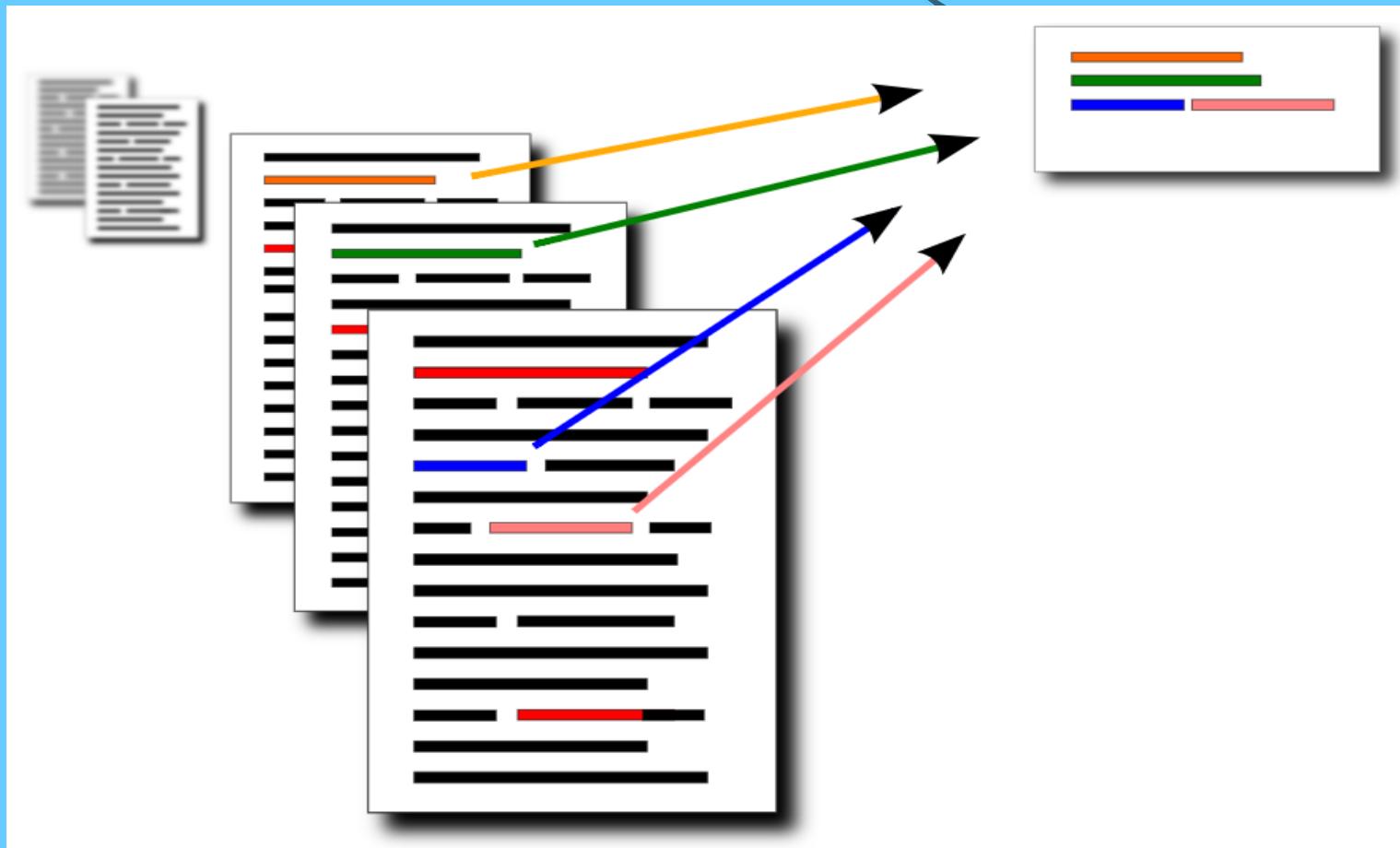
Hãy phân loại các đánh giá sau là tích cực hay tiêu cực:

1. Cán bộ trực ban giao tiếp rất khó chịu, giọng điệu hằn học. Gần 14h chiều ngày thứ ba nhưng bàn trực ban không có ai, 3-4 người dân phải ngồi đợi.
2. Tại địa chỉ số 215/36C, bãi xe Nam, Nguyễn Xí, Phường 13, Quận Bình Thạnh, nhà lấn chiếm hết hẻm làm bãi giữ xe, xây nhà và trổ cửa trái phép lấn ra sông Vầm Thuật.
3. Việc đốt (nhóm) bếp than vào mỗi sáng sớm từ 3 đến 4 giờ sáng của hộ gia đình 144/10 đường Lê Văn Thọ, phường 11, quận Gò Vấp rất độc hại. Gây ảnh hưởng âm thầm đến sức khỏe của rất nhiều người sinh sống xung quanh.
4. Về việc gần tết, các cán bộ không hề lơ là trong công việc và luôn có mặt ở cơ quan trong giờ làm việc hành chánh để hỗ trợ người dân.
5. 133 Nguyễn Tiểu La, Phường 5 Quận 10, nhà cho thuê để các bao xi măng, bao cát, đá lấn chiếm lòng lề đường mấy năm nay. Người dân có liên hệ UBND Phường nhưng không được giải quyết triệt để.



1. Tiêu cực: Cán bộ trực ban không được hỗ trợ và cách hành xử của họ rất khó chịu.
2. Tiêu cực: Ngôi nhà lấn chiếm hẻm, làm bãi giữ xe, xây nhà và trổ cửa trái phép, lấn ra sông Vầm Thuật.
3. Tiêu cực: Hộ gia đình đốt (nhóm) bếp than vào mỗi sáng sớm từ 3 đến 4 giờ sáng, gây ảnh hưởng đến sức khỏe của nhiều người.
4. Tích cực: Các cán bộ không lơ là trong công việc, luôn có mặt ở cơ quan và hỗ trợ người dân trong giờ làm việc hành chánh.
5. Tiêu cực: Nhà cho thuê bị lấn chiếm lòng lề đường mấy năm nay bởi các bao xi măng, bao cát, đá và việc này chưa được giải quyết triệt để bởi UBND Phường.

#### 4. Text summarization:



# ChatGPT: text summarization

DU

Hãy tóm tắt Nghị định số 72/2013/NĐ-CP ngày 15/7/2013 về Quản lý, cung cấp, sử dụng dịch vụ Internet và thông tin trên mạng;

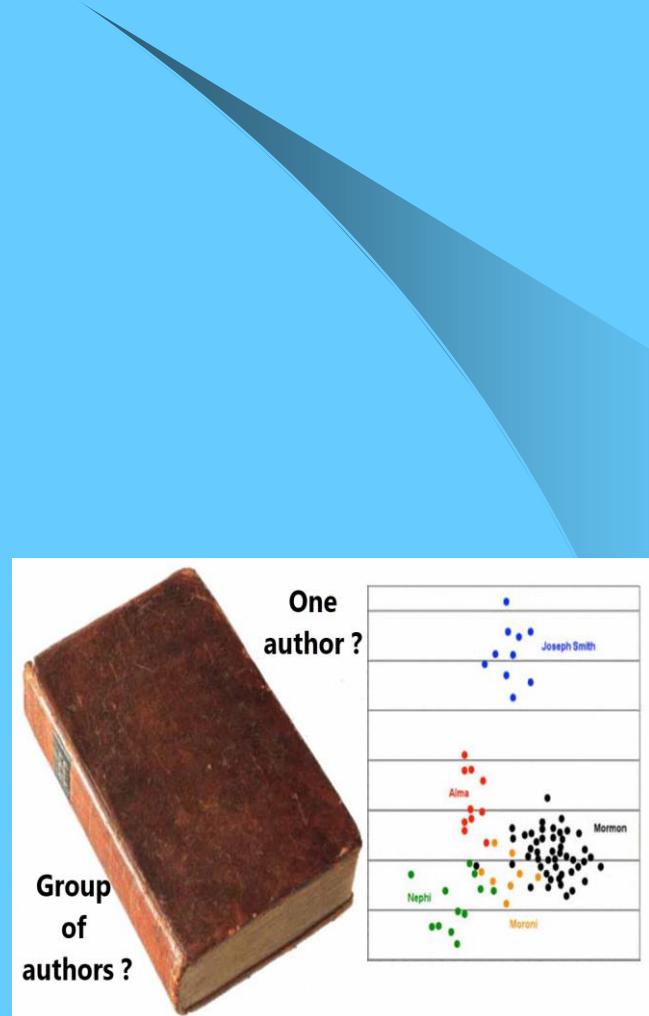
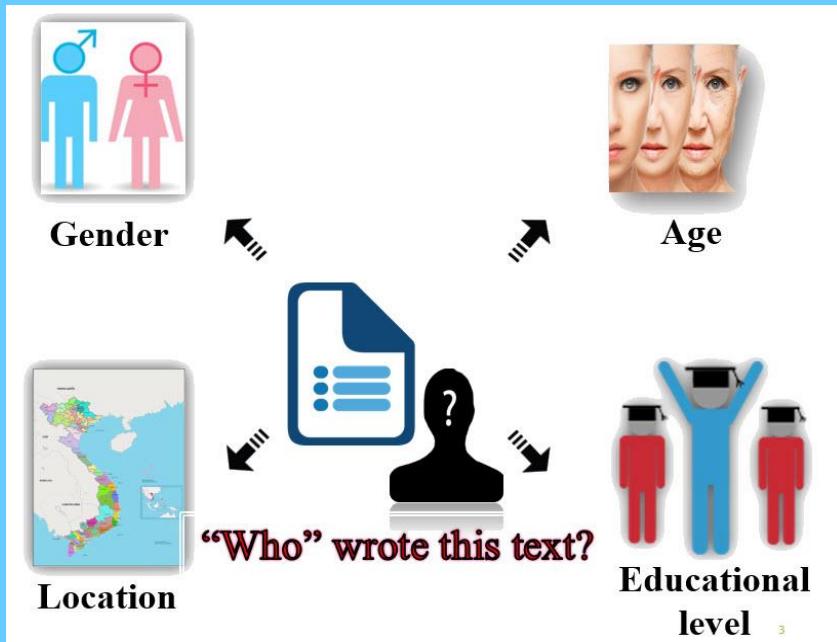


Nghị định số 72/2013/NĐ-CP được ban hành ngày 15/7/2013 về Quản lý, cung cấp, sử dụng dịch vụ Internet và thông tin trên mạng có các điểm chính như sau:

1. Đối tượng áp dụng: các tổ chức, cá nhân cung cấp, sử dụng dịch vụ Internet và thông tin trên mạng.
2. Quy định về việc sử dụng Internet và thông tin trên mạng, bao gồm quản lý, bảo mật, và đảm bảo an toàn thông tin.
3. Quy định về cung cấp dịch vụ Internet, quyền và nghĩa vụ của nhà cung cấp, cơ chế quản lý và giám sát hoạt động của nhà cung cấp dịch vụ Internet.
4. Quy định về trách nhiệm của các tổ chức, cá nhân sử dụng dịch vụ Internet và thông tin trên mạng, bao gồm việc bảo vệ quyền lợi của người dùng và trách nhiệm trong việc cung cấp thông tin trên mạng.
5. Quy định về kiểm tra, giám sát hoạt động sử dụng Internet và thông tin trên mạng.
6. Quy định về xử lý các hành vi vi phạm pháp luật khi sử dụng Internet và thông tin trên mạng.

Nghị định này nhằm đảm bảo an toàn, bảo mật và phát triển bền vững của dịch vụ Internet và thông tin trên mạng, đồng thời đảm bảo quyền lợi và trách nhiệm của các tổ chức, cá nhân sử dụng dịch vụ này.

## 5. Text stylometry:



# Văn phong khác biệt tố cáo vụ ngộ sát, làm giả thư tuyệt mệnh

**Qua phân tích cách dùng từ "and", "but", "hopefully", "truly" trong thư tuyệt mệnh, các chuyên gia tại Mỹ xác định nạn nhân không phải là người viết.**

- Thói quen lạ của chú chó tố cáo ông chủ vứt xác bốn cô gái bán dâm

Vào buổi sáng năm 1992, Khoa cấp cứu một bệnh viện tại Mỹ nhận được cuộc gọi khẩn cấp từ người sống tại căn hộ ở Bắc Carolina. Khi đến nơi, các nhân viên y tế thấy một thanh niên đã tử vong.

Nạn nhân được xác định là Michael Hunter, 23 tuổi, vừa tốt nghiệp đại học và đang làm lập trình viên. Bạn cùng phòng khai với cảnh sát rằng sáng hôm ấy, khi đánh thức Michael Hunter dậy để đi làm thì thấy anh ta bất tỉnh từ bao giờ.

Michael Hunter không có thương tích khả nghi nào trên cơ thể. Xét nghiệm máu cho kết quả dương tính với một loại thuốc gây tê với nồng độ gây chết người. Thông thường, loại thuốc này được sử dụng trong một số trường hợp khẩn cấp để làm ổn định nhịp tim. Tuy nhiên, nhân viên y tế khẳng định khi đến nơi thì thấy nạn nhân đã tử vong và họ không hề tiêm bất cứ thuốc gì.

Cái chết của Michael Hunter làm gia đình anh suy sụp. Cha của anh vì quá đau buồn đã rơi vào cơn trầm cảm kéo dài và tự tử sau đó.



# FBI Profiler Says Linguistic Work Was Pivotal In Capture Of Unabomber

August 22, 2017 · 12:18 PM ET

Heard on Fresh Air

DAVE DAVIES

FRESH AIR



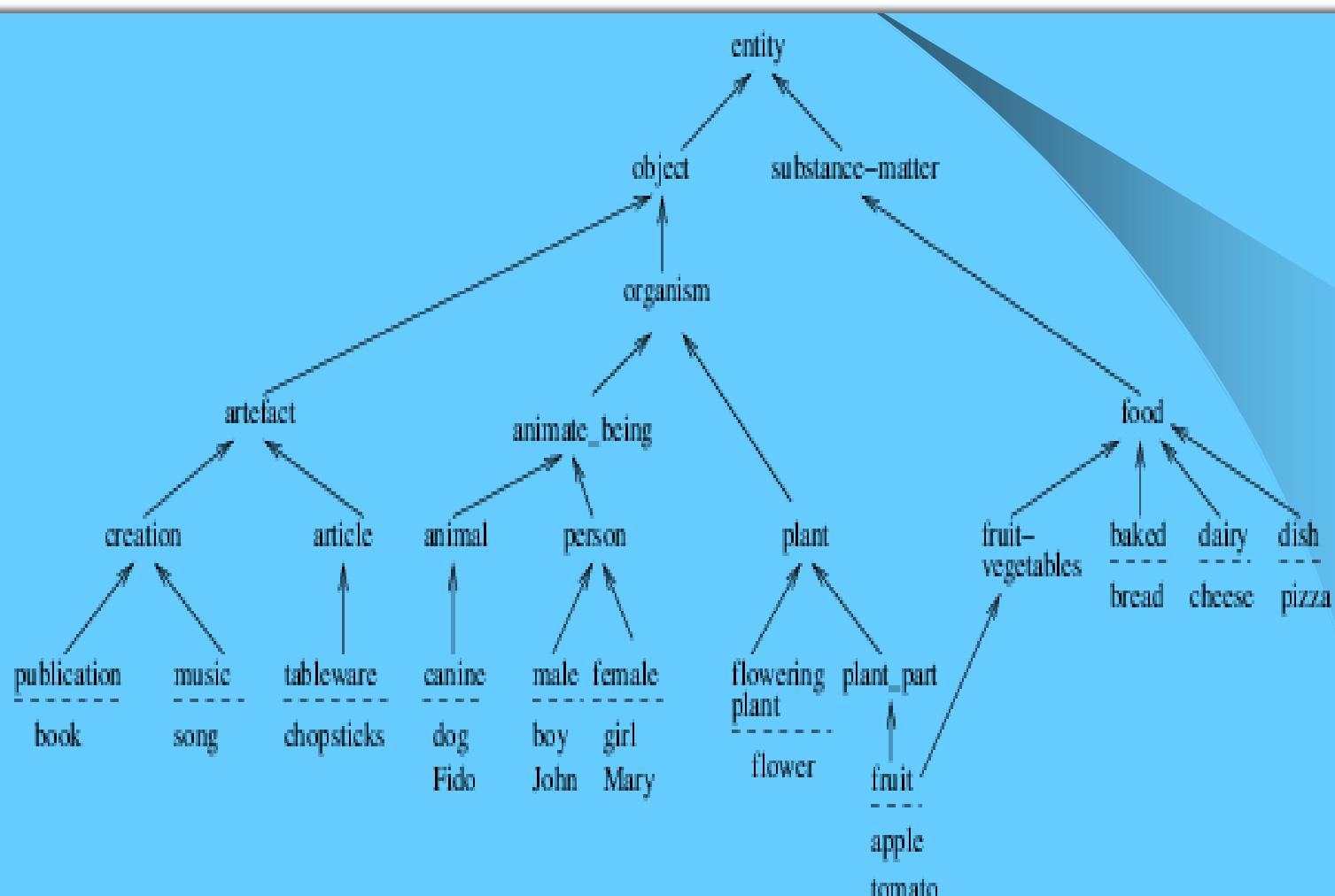
Ted Kaczynski is flanked by federal agents as he is led to court. Kaczynski is now serving a life sentence in prison for the

his victims. In 1995, he sent a sprawling, 35,000-word "manifesto" to *The New York Times* and *The Washington Post*, in which he explained why he believed technology to be evil and how society should disband the technological system and live in agrarian tribes.

Ex-Math-Prof.  
UC Berkeley

Fitzgerald says the Unabomber's writings were a "pivotal factor" in cracking the case. He and his colleagues used them to help pinpoint the age and geographic origin of their suspect — evidence that helped lead to the April 6, 1996, arrest of Ted Kaczynski.

## 6. Text similarity:



# Plagiarism detection

noplag

Title: Health Vision(1)  
Author: Aleks B

100%  
Similarity  
43 Matches  
en Language

You have not seen your eye doctor for more than a year.

What eye problems your eye doctor is looking for?

- ? Nearsightedness, farsightedness or astigmatism. These conditions are corrected with eyeglasses, contact lenses or surgery.
- ? Amblyopia and strabismus. Amblyopia occurs when eyes are misaligned. Strabismus is another word for crossed eyes.
- ? Focusing problems and ability of your eyes to work together.
- ? Any problems with eye tearing.
- ? Eye diseases such as glaucoma and diabetic retinopathy which have no clear symptoms at early stages. In most cases, early detection can reduce risk for vision loss.
- ? Age-related conditions. For example, cataracts occur mostly at the age of 65 and older.

What can you do to protect your eyes?

- ? Have a healthy diet, rich in fruits and vegetables.
- ? Take care of your health in general.
- ? Maintain a healthy weight.
- ? Quit smoking.
- ? Remember to give your eyes a rest when working at the computer.
- ? Do not forget to blink.
- ? Keep your eyes safe when playing sports or doing any potentially eye-dangerous activity.
- ? Protect your eyes from ultraviolet rays with sunglasses.
- ? Know your family's eye health history. Many eye diseases and conditions are hereditary.
- ? Visit your eye doctor once a year. Conducting regular eye exams will help preserve your vision and reduce risk of serious eye and vision problems.

Originality report Powered by Noplag.com

Page 1

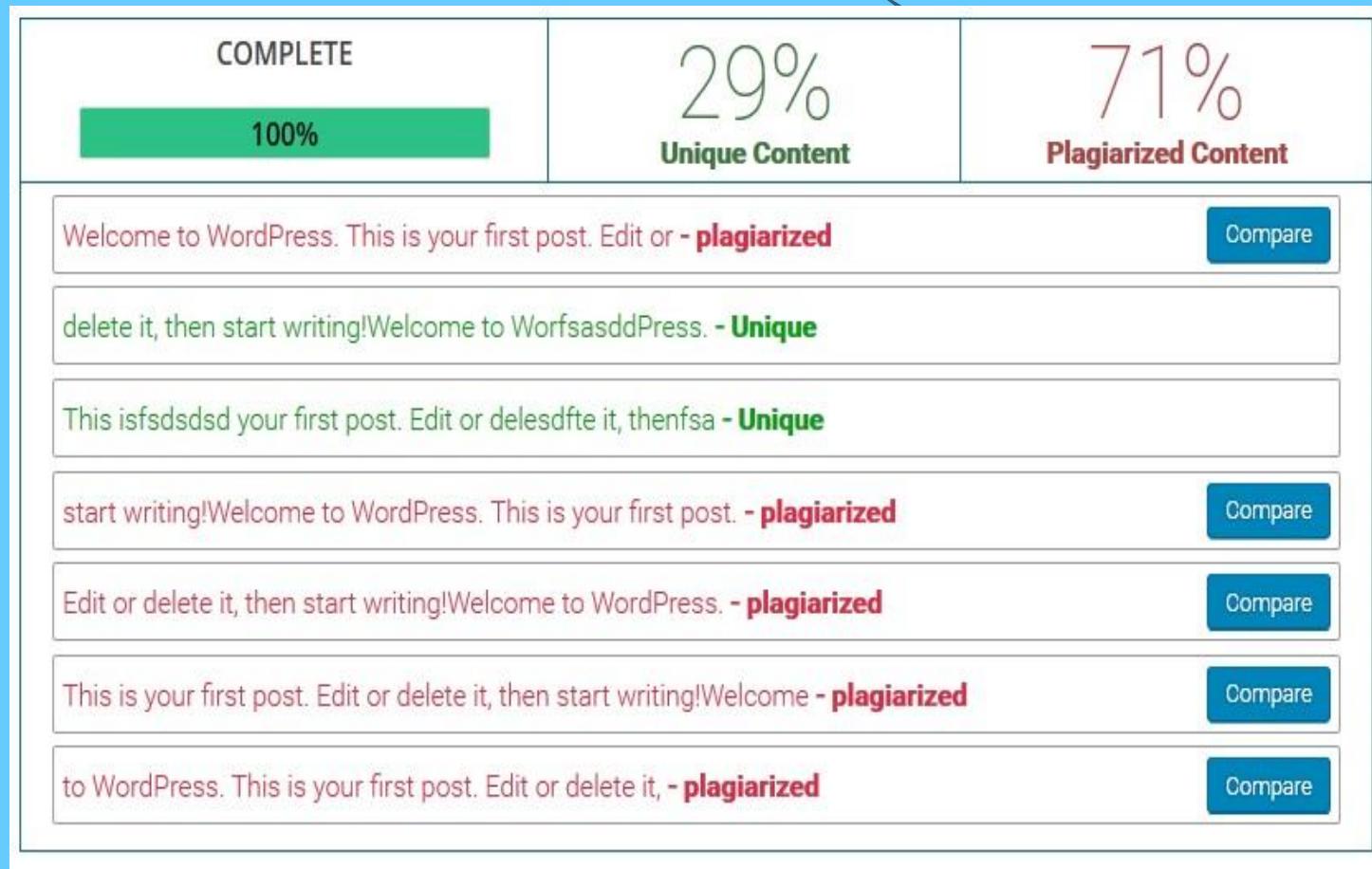
Match Overview

Rank	Type	Similarity
1	My Library	W 364   100%
2	Web	W 345   95.40%
3	Web	W 345   95.40%
4	Web	W 345   95.40%
5	Web	W 345   95.40%
6	Web	W 39   10.93%
7	Web	W 33   9.68%
8	Web	W 27   7.91%
9	Web	W 23   6.69%

ID 468042 Checked on 04 Nov 2016 5:42 PM

Words: 364 Pages: 1/1

# Cross-Lingual Plagiarism detection



## 7. Text readability:

### Earthquake in Indonesia – level 1



02-10-2018 07:00

Level 1

Level 2

Level 3

Sulawesi is an island in Indonesia. An **earthquake** hits near it. The earthquake makes a **tsunami**. It is 3 metres tall.

The tsunami moves into two cities. Around 600,000 people live there. More than 832 people die. Hospitals, hotels, a shopping centre, and thousands of homes are **destroyed**.

Difficult words: **earthquake** (when the ground moves), **tsunami** (a big wave started by an earthquake), **destroy** (break completely).

## Earthquake in Indonesia – level 2



02-10-2018 07:00

Level 1

Level 2

Level 3

A 7.5-magnitude earthquake hit near the Indonesian island of Sulawesi which triggered a 3-metre tsunami that smashed into two cities on the coast. These cities are home to 600,000 people.

The tsunami killed more than 832 people and destroyed hospitals, hotels, a shopping centre, and thousands of homes. The event affected the lives of as many as 1.6 million people.

Difficult words: **magnitude** (the size of power of something), **trigger** (start suddenly), **smash** (move into with a lot of force).

## Earthquake in Indonesia – level 3



02-10-2018 07:00

Level 1

Level 2

Level 3

A 7.5-magnitude earthquake hit near the Indonesian island of Sulawesi, triggering a 3-metre tsunami, which smashed into two cities on the coast.

Palu and Donggala are the cities affected the worst, and they are home to over 600,000 people. At least 832 people have been confirmed dead, thousands of homes collapsed, along with hospitals, hotels, and a shopping centre. The disaster affected as many as 1.6 million people, according to Red Cross estimates.

Difficult words: trigger (start), estimate (a careful guess based on data).

## **Linguistic features of Text readability**

Word popularity: word usage frequency

Syntactic structure: complexity of parsing tree

Text organization: text coherence

Text readability <> comprehensibility

Writer (encoder) <> Reader (decoder)



Ex: A top-3,000 wordlist in English has been used in all definitions/explanations in the Oxford OALD8,

e.g. **phil·an·throp·ist** /fɪ'lænθrəpɪst/ noun a rich person who helps the poor and those in need, especially by giving money •nhà từ thiện, mạnh thường quân

Whilst, in an existing Vietnamese dictionary: the definition of the word “đường” (sugar) is “một hợp chất kết tinh...” (“hợp chất” = compound, “kết tinh” = crystallize”).

Ex: “tòa” = “kiến trúc đơn nguyên trong xây dựng”

Should not use difficult words: “gà qué” (35.216), “con ngóe” (23.670), ... in grade-1 textbooks.

# MS word\proof reading: available for English

HubSpot Blogs - Marketing scanned on 14 Apr 2015 | Run new scan | New folder | More ▾

Summary Clear Language Links Spelling Bad Language Good Language Discovery Activity Discussions

## HUBSPOT BLOGS - MARKETING Clarity Grader Report

Url Scanned: <http://blog.hubspot.com/marketing>

The Clarity Grader report analyzes this site for **clear, transparent** language.  
We also check for **consistent language** using customizable bad and good language dictionaries.

---

1 PAGES SCANNED ON 14 APRIL 2015

[Tweet Report](#) [Email Report](#) [PDF this Report](#)



### Clear Language

Long Sentences 71 Sentences <b>25.27%</b>	Average Sentence Length <b>14</b>	Passive Language 9 Sentences <b>3.20%</b>	Readability <b>62</b>
---	--------------------------------------	---	--------------------------

**Aim for 5% or lower**

Long sentences exceed 20 words. At 25.27% your content is 5.1 times the recommended level of 5%. The message is likely buried in complex statements and run on sentences. Split the long sentences or use lists.

**Aim for 10 or lower**

The average sentence length is fair at 14. For web copy you should aim for 10 or less. You may be burying certain key messages.

**Aim for 5% or lower**

The passive voice % is good at 3.20%, Well done! Your text is punchy and active. This means readers can easily absorb your message and follow instructions.

**Aim for at least 60**

Great. Your **readability** score is above 60. Your message is clear and readers can easily follow instructional text.

## 8. Text translation:

### 8a. Machine Translation

The screenshot shows the Al Jazeera English website interface. At the top, there's a banner with Arabic text "الجزيرة نت" and English text "Speech and ... of Contents Book Schedule University o... uter Science James Mart... Home Page The Daily Camera The New York...Multimedia". Below the banner, there are two main news headlines:

- أشرتك الآن** (Buy it now) featuring a large image of a bread roll.
- انت في التقارير تفتتح الخط الأرضي والمتاحل للنمرة** (You are in the reports, opening the ground line and the available for the nمرة) featuring an image of several men in suits.

Below these, there are several smaller news items with both English and Arabic summaries. One prominent article is about the killing of Palestinians during an Israeli air raid in Gaza. Other articles include stories about Bashir meeting Fraser, Rumsfeld and Cheney insisting on keeping American forces in Iraq, and a suicide attack in Afghanistan.

#### Killing Palestinians and wounding nine in the raids Sector

Nine Palestinians were wounded among civilians in an Israeli air raid in the neighborhood result in the Gaza Strip. This comes immediately after the killing of two prominent Al-Aqsa Martyrs Brigades in the Israeli occupying forces carried out air and infantry forces in the Balata camp in the West Bank.



#### Bashir meets Fraser, the Security Council will not impose forces Darfur

Is scheduled to meet with Sudanese President Omar al-Bashir Jenday Fraser Assistant Minister for Foreign Affairs of the American attempt to persuade officials in Khartoum, Sudanese Darfur deployment of the nationalities. For his part, US Ambassador to the United Nations that it has no intention of the Security Council to impose its forces in the province.



#### Rumsfield and Cheney insist on keeping the American forces in Iraq

Called American Defense Minister Donald Rumsfield Americans to show patience on Iraq. I take Vice President Dick Cheney calls Democrats withdrawal of American forces from Iraq link and the possibility of early withdrawal of attacks inside the United States.



#### Killing civilians and wounding officer suicide attack in Afghanistan

The international force to help establish security (ISAF) killed civilians and the wounding of an officer in an attack against Afghan forces convoy south Atlantic Afghanistan. In the capital Kabul, a hand grenade exploded at the passage of manufacture French patrol was not reported injuries or damage.



# ChatGPT translation

DD

Hãy dịch câu sau sang tiếng Anh, Pháp, Hoa, Nhật, Hàn, Đức, Nga, Latin và Quốc tế ngữ: "Chúng tôi đang học một ngôn ngữ"



- Tiếng Anh: "We are learning a language"
- Tiếng Pháp: "Nous apprenons une langue"
- Tiếng Hoa: "我们正在学习一种语言" (Wǒmen zhèngzài xuéxí yī zhǒng yǔyán)
- Tiếng Nhật: "私たちは言語を学んでいます" (Watashitachi wa gengo o manande imasu)
- Tiếng Hàn: "우리는 언어를 배우고 있습니다" (Urineun eoneoreul baeugo issseubnida)
- Tiếng Đức: "Wir lernen eine Sprache"
- Tiếng Nga: "Мы изучаем язык" (My izuchayem yazyk)      учим
- Tiếng Latin: "Nos discimus linguam"
- Quốc tế ngữ: "Emang volapükik"                                Ni lernas lingvon

# ChatGPT translation

DD

Hãy dịch câu sau sang tiếng Anh, Pháp, Hoa, Nhật, Hàn, Đức, Nga, Latin và Esperanto: "Chúng tôi đang học một ngôn ngữ"



- Tiếng Anh: "We are learning a language"
- Tiếng Pháp: "Nous apprenons une langue"
- Tiếng Hoa: "我们正在学习一门语言" (Wǒmen zhèngzài xuéxí yī mén yǔyán)
- Tiếng Nhật: "私たちは言語を学んでいます" (Watashitachi wa gengo o manande imasu)
- Tiếng Hàn: "우리는 언어를 배우고 있습니다" (Urineun eoneoreul baeugo issseubnida)
- Tiếng Đức: "Wir lernen eine Sprache"
- Tiếng Nga: "Мы изучаем язык" (My izuchayem yazyk)
- Tiếng Latin: "Nos discimus linguam"
- Esperanto: "Ni lernas lingvon"



# ChatGPT: cannot process Nôm-scripts

- “nǎm” (5): 卯 = 南 (sound /nán/ “nam”) + 五 (meaning: *ngũ*)
- “nǎm” (year): 酉 = 南 (sound /nán/) + 年 (meaning: *niên*)
- “中” => “đúng, trong, trung, trúng, truồng”
- 穀 = “trăm/trặn” ;



Hàn  
Thuyên



Alexandre de  
Rhodes



季 群 底 固 故 膺

Trời còn để có hôm nay,

散 霜 頭 午 援 遽 钟 季

Tan sương đầu **ngõ**, vén mây giữa trời.

nomfoundation.org/nom-tools/Nom-Lookup-Tool/Nom-Lookup-Tool?uiLang=vn



會 保 存 遺 產 哺

English

Thông tin về Hội ▼ Chữ Nôm ▼ Sách Nôm ▼ Từ điển Nôm ▼ Dự án Nôm ▼ Phòng chữ Nôm ▼ Liên hệ

Another Nôm Lookup Tool  
based on Unicode

- ◎ Quốc Ngữ hoặc Hán-Nôm  Mã Unicode hoặc TCVN (dùng hệ hex)  Tiếng anh  
 Bắc kinh  Quảng đông  Thương hiệt  
 Bộ thủ.Tổng nét

午

GO

TRA CỨU CHỮ NÔM

Tra cứu chữ Nôm

Câu chuyện về tra cứu Nôm

Quốc Ngữ	Hán-Nôm	Context	Ref.	Tiếng anh
ngõ	午	cửa ngõ	btcn	gateway
ngọ	午	giờ ngọ	vhn	midday, noon
ngọ	午	ngó ý	btcn	to express a wish

# Auto translation Hán-Nôm sang chữ Quốc ngữ

## Truyện Kiều

慕辭沖 培 得 些

Trăm năm trong cõi người ta

字 才 字 命 簿 羅 恕 饒

Chữ tài chữ mệnh khéo là ghét nhau

浪辭嘉靖朝明

Rằng: Năm Gia Tinh triều Minh

眾方滂朗台京凭傍

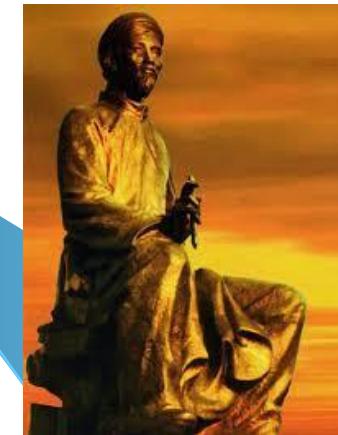
Bốn phương phảng lặng hai kinh vững vàng

固茹員外戶王

Có nhà viên ngoại họ Vương

家資擬拱常常塙中

Gia tư nghỉ cũng thường thường bậc trung



Nguyen Du  
(1766-1820)

- *nghĩ* (think)
- *nghỉ* (he)



# Nghiệm thu hệ thống tự động chuyển văn bản chữ Nôm sang chữ Quốc ngữ

Xem nhiều

① 14/02/2023 - 19:33

Chia sẻ **Ý KIẾN BẠN ĐỌC(2)**



PNO - Các nhà khoa học của TPHCM đã xây dựng thành công hệ thống tự động chuyển văn bản chữ Nôm sang chữ Quốc ngữ.



- Chữ quốc ngữ 130 năm thăng trầm
- Lại bàn về việc cải tiến chữ quốc ngữ
- Bài 1: "Ông tổ thơ Nôm" đúng nhìn "cha đẻ chữ Quốc ngữ"

GSTS Võ Văn Sen

Công việc này có ý nghĩa rất lớn-Phải nói là hết sức vĩ đại!!! Cảm tạ nhóm tác giả!!!

Trả lời 1 thích 0 không thích

Nguyễn Ngọc Lan

Khâm phục  
Tâm rất sáng, công rất lớn...

Trả lời 1 thích 0 không thích



Hội đồng nghiệm thu nhiệm vụ khoa học công nghệ

Sở Khoa học và Công nghệ TPHCM vừa tổ chức Hội đồng nghiệm thu nhiệm vụ khoa học và công nghệ "Xây dựng hệ thống chuyển tự từ văn bản chữ Nôm sang chữ Quốc ngữ". Đây là nhiệm



học sinh thiếu sóch  
giáo khoa mỹ thuật



Tư vấn nguyện vọng  
tuyển sinh lớp 10:  
Đừng gieo cǎng thẳng  
đôi bên



"Có thể học được kỹ  
năng kiến tạo hạnh  
phúc không?"

Tuyển sinh



Hà Nội chốt lịch thi vào  
lớp 10 THPT, sớm một  
tuần so với năm ngoái



Đầu tháng 3 sẽ công  
bố mẫu đề thi tốt  
nghiệp THPT



Hà Nội chốt thi 3 môn  
vào lớp 10 công lập

HÀN - NÔM

现世为一师當來作佛租  
有禪有淨土橫如帶角虎

QUỐC NGỮ

hiện thế vi nhất sư đương lai tác phật tō  
hữu thiên hữu tịnh thổ hoành như dải giác hổ



21 / 500



X



# 宁 喃

chữ Nôm

fit@hcmus



## CLC - Chuyển tự chữ Nôm

VĂN BẢN

TÀI LIỆU

HÌNH ẢNH

LỊCH SỬ

ĐÃ LƯU

 Vietnamese English

HÁN - NÔM



QUỐC NGỮ

人有憂愁優  
點大nhân hữu  
ưu sầu ưu  
điểm đạiDịch nghĩa  
(thử nghiệm):  
người biết lo  
âu ưu điểm lớn

0 / 500



Công trình này được thực hiện bởi một số giảng



# Voice translation

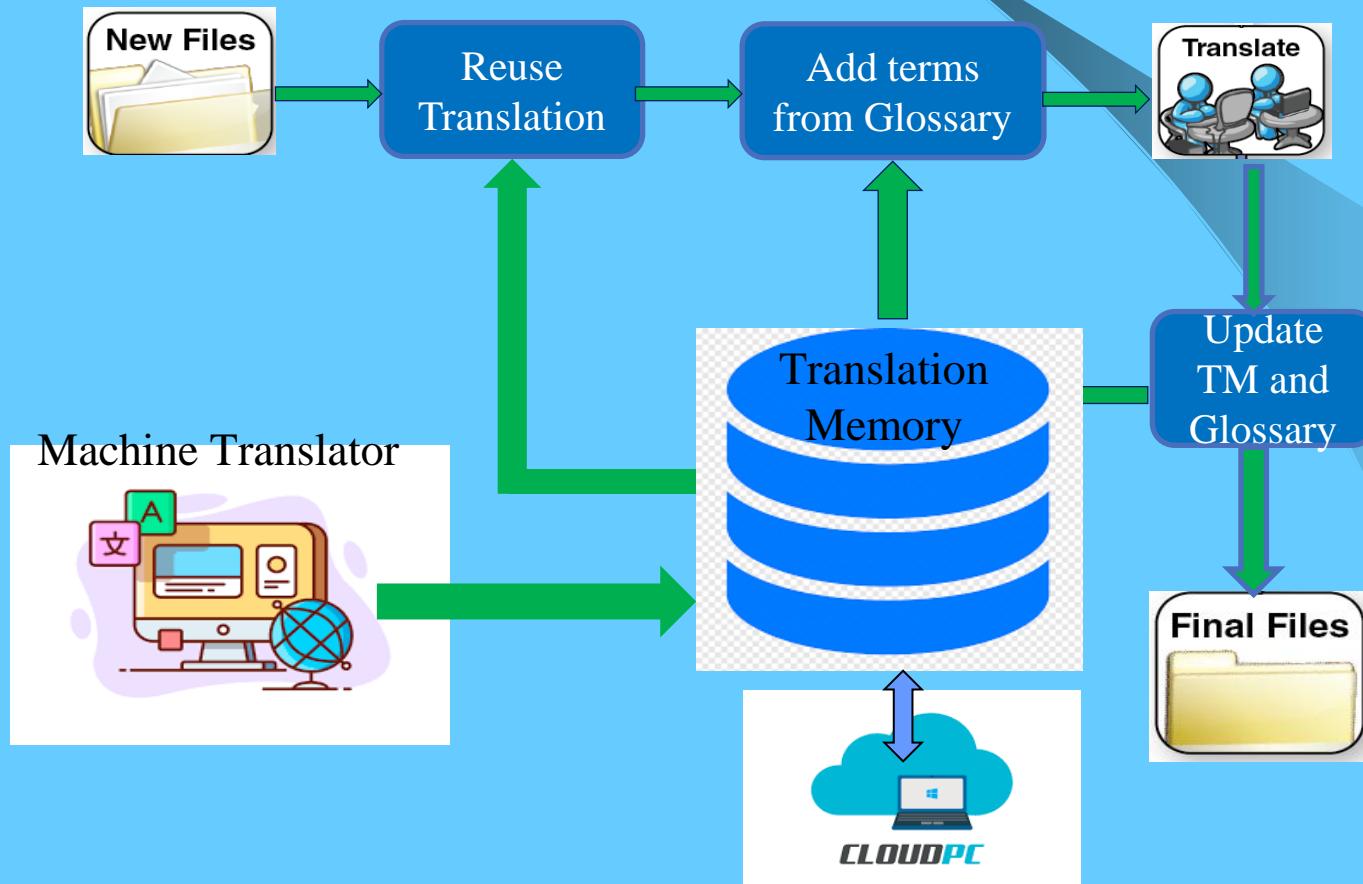


# Image translation



## 8a. CAT

# Computer Assisted Translation Translation Memory, Glossary



# Giới thiệu bộ nhớ dịch (Translation Memory)

## Dóng hàng đoạn:

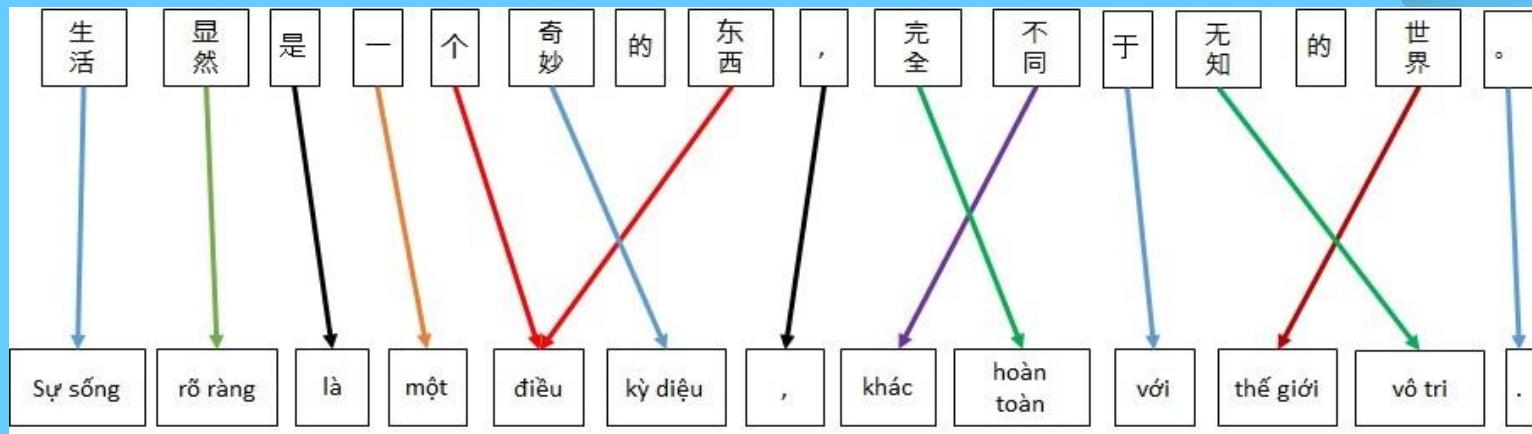
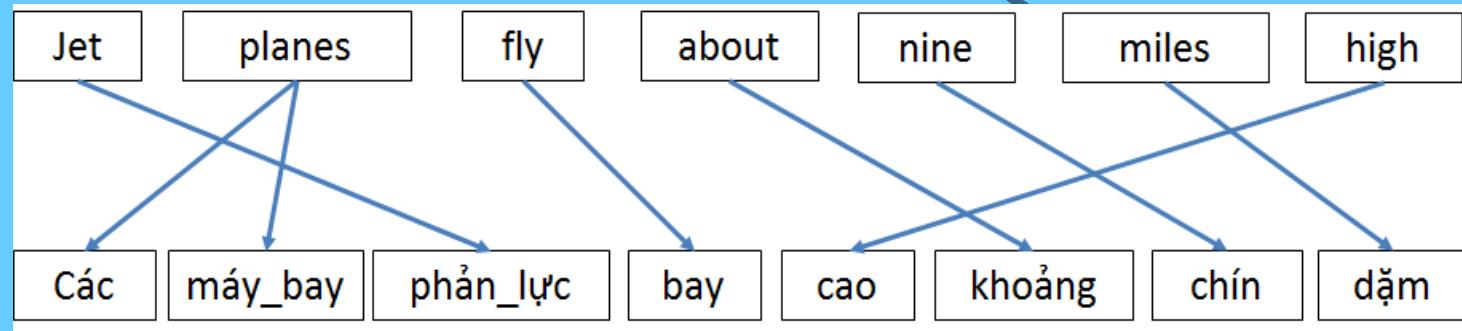
Helicopters can rise straight up into the air and can go straight down. They can stand still in the air. Helicopters do not have wings. A huge whirling propeller, called a rotor, on top of a helicopter provides the lift.

Máy bay trực thăng có thể lên thẳng trên không và đáp thẳng xuống đất. Chúng có thể đứng yên trên không. Máy bay trực thăng không có cánh, một cánh quạt lớn gọi là chong chóng trên đầu chiếc máy bay cung cấp sức nâng.

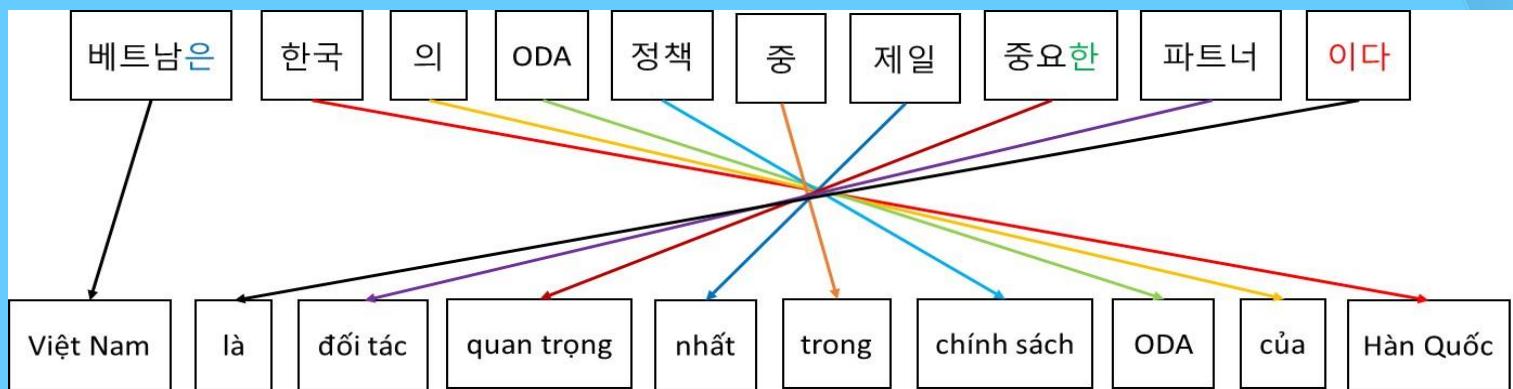
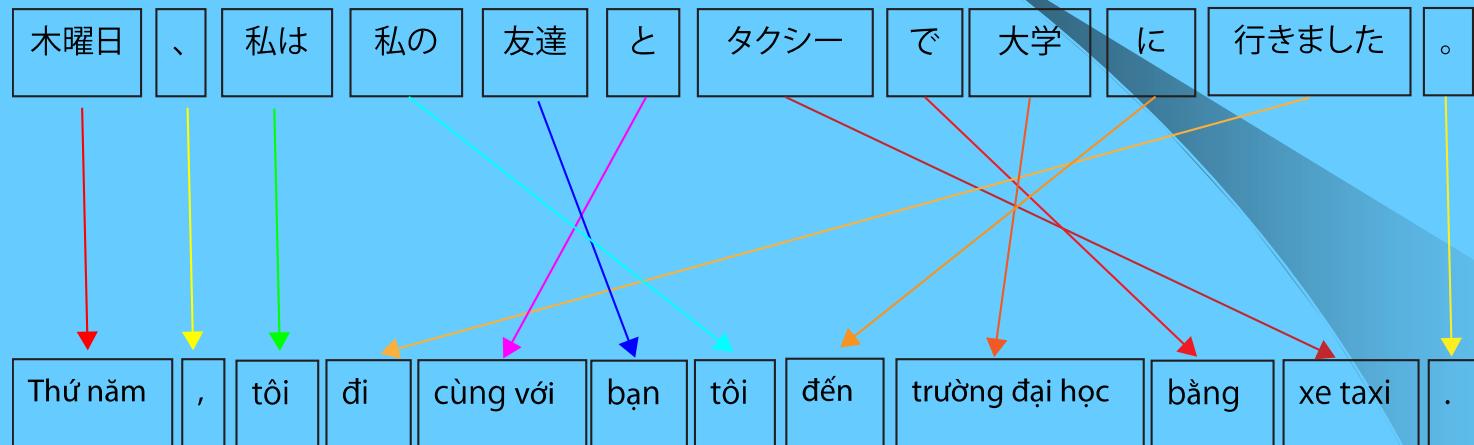
## Dóng hàng câu (cặp câu song song):

- \* Helicopters can rise straight up into the air and can go straight down.  
+ Máy bay trực thăng có thể lên thẳng trên không và đáp thẳng xuống đất.
- \* They can stand still in the air.  
+ Chúng có thể đứng yên trên không.
- \* Helicopters do not have wings.  
+ Máy bay trực thăng không có cánh.

## 9. Parallel text processing: Contrastive Linguistics



# Parallel Corpora



Search Word

Key search		xảy ra	
<input type="radio"/> Match cases	<input checked="" type="radio"/> Morphological cases	<input type="radio"/> English	<input checked="" type="radio"/> Vietnamese
		Search	Find 510 results
Left	Key	Right	
...gional peacekeeping efforts, something we would like to see	happen	as Vietnam begins its campaign for a seat on the UN Security .	
...kier investment destination because unanticipated problems	occur	more frequently here than elsewhere.	
...government of Vietnam to make sure that child selling never	occurs	.	
What will	happen	to small, subsistence farmers in their own local markets in dev.	
...everyone is comfortable with the kinds of possible changes to	come	.	
Upon the	occurrence	of the Indian Ocean tsunami in late 2004, the Chinese govern.	
Most shops in areas where the violence	occurred	remained closed as of Monday night.	
... chief told media on Tuesday that a civil war was unlikely to	occur	.	
... called her younger brother in Urumqi, " A lot of things have	happened	, and we all know something might happen in Urumqi tomorrow	
...ng as an inseparable part of China and believes that what is	going	on there is exclusively an internal affair for China ", the ministr.	

Left		Key	Right
...đang đối rủi ro và đắt đỏ hơn vì những vấn đề bất ngờ ở đây	xảy ra	thường xuyên hơn những nơi khác.	
... Việt Nam để đảm bảo rằng mua bán trẻ em không bao giờ	xảy ra	.	
Điều gì sẽ	xảy ra	với những người tiêu nông, làm chỉ đủ ăn tại thị trường của c.	
...để mọi người đều có thể thoải mái về những thay đổi có thể	xảy ra	.	
Khi	xảy ra	sóng thần ở Ấn Độ Dương vào cuối năm 2004, chính phủ và	
Hầu hết các cửa hàng trong các khu vực mà bạo lực đã	xảy ra	vẫn tiếp tục đóng cửa giống như đêm hôm thứ hai.	
... với báo chí vào ngày thứ ba là một cuộc nội chiến đã không	xảy ra	.	
...khi bà gọi cho anh trai của bà ở Urumqi, " Rất nhiều điều đã	xảy ra	, và tất cả chúng ta biết điều gì có thể xảy ra ở Urumqi vào đ.	
...hông thể tách rời của Trung Quốc và tin rằng những gì đang	xảy ra	chi là việc nội bộ của Trung Quốc ", Bộ ngoại giao cho biết trc	
...lbo, cho biết chính phủ Trung Quốc đã rất cởi mở vào ngày	xảy ra	vụ xô xát.	

Search Word

Found total 217 results.			
Left	Key	Right	
He	wears	a ring on his middle finger.	
He	wears	an identity disc round his neck.	
He	wears	clean socks every day.	
He	wears	his brother 's cast - offs.	
He	wore	a cap with flaps to cover his ears.	
He	wore	a gold chain round his neck.	
He	wore	a hat, gloves and and overcoat.	
He	wore	a thick overcoat as a protection against the bitter cold.	
He	wore	his robes as a token of office.	
He	wore	his chubbliest clothes to the party. <i>ba ba no sona of... III</i>	

Vietnamese			
Left	Key	Right	
Ông ta	đeo	một chiếc nhẫn ở ngón giữa.	
Nó	đeo	một thẻ tròn nhận dạng nơi cổ.	
Anh ta	mang	bít tất sạch hằng ngày.	
Nó	mặc quần áo	thứa của anh nó.	
Anh ta	đội	mũ có vạt che tai.	
Anh ấy	đeo	một sợi dây chuyền vàng trên cổ.	
Ông ấy	đội	một chiếc mũ, đị đội găng tay và mặc một cái áo khoác.	
Ông ta đã	mặc	một chiếc áo khoác dày để chống lại cái lạnh cắt da.	
Ông ta	mặc	chiếc áo choàng như là một biểu tượng chức vụ của ông	
III	mặc	những quần áo đặc biệt nhất đến địa điểm hoành tráng nhất	

Untitled - Parallel Corpus Processor

File Edit View Statistic Help

	A	HUM
1	the jury said , " considering the widespread interest in the election , the number of	voters and the size of this city "
2	It recommended that Fulton	legislators act " to have these laws :
3	The	jurors said they realize " a prop
4	he future Fulton County should receive some portion of these available funds " , the	jurors said .
5	" Failure to do this will continue to place a disproportionate burden " on Fulton	taxpayers .
6	on ordinary 's court which has been under fire for its practices in the appointment of	appraisers , guardians and administ
7	" These actions should serve to protect in fact and in effect the court 's	wards from undue costs and its
8	but it added that " there should be periodic surveillance of the pricing practices of the	concessionaires for the purpose of keepir
9	On other matters , the jury recommended that : Four additional	deputies be employed at the Fulto
10	Fulton	legislators " work with city officials
11	Mayor William B. Hartsfield filed suit for divorce from his	wife , Pearl Williams Hartsfie
12	They have a	son , William Berry Jr. , and
13	The petition listed the	Attorneys for the mayor said that a
14	It listed his	mayor 's occupation as " attorne
15	Henry L. Bowden was listed on the petition as the	wife 's age as 74 and place of
16	Hartsfield has been	mayor 's attorney .
17	The	mayor of Atlanta , with exceptio
18	He will be succeeded by Ivan Allen Jr. , who became a	mayor 's present term of office
19	Georgia	candidate in the Sept. 13 primary a
20	Robert Snodgrass , state GOP	Republicans are getting strong encour
21	. 8 in Savannah at which newly elected Texas Sen. John Tower will be the featured	chairman , said a meeting held Tue
22	In the Blue Ridge meeting , the audience was warned that entering a	speaker .
23	Despite the warning , there was a unanimous vote to enter a	candidate for governor would force
24	The largest hurdle the	candidate , according to Repuplican
25	Sam Caldwell , State Highway Department public relations	Republicans would have to face is a s
26	He will be succeeded by Rob Ledford of Gainesville , who has been an	director , resigned Tuesday to wo
27	hen the gubernatorial campaign starts , Caldwell is expected to become a campaign	assistant more than three years .
28		coordinator for Byrd .

Search Word

Language  Korean  Vietnamese

Key search

Match cases  Morphological cases  Phrase

Korean

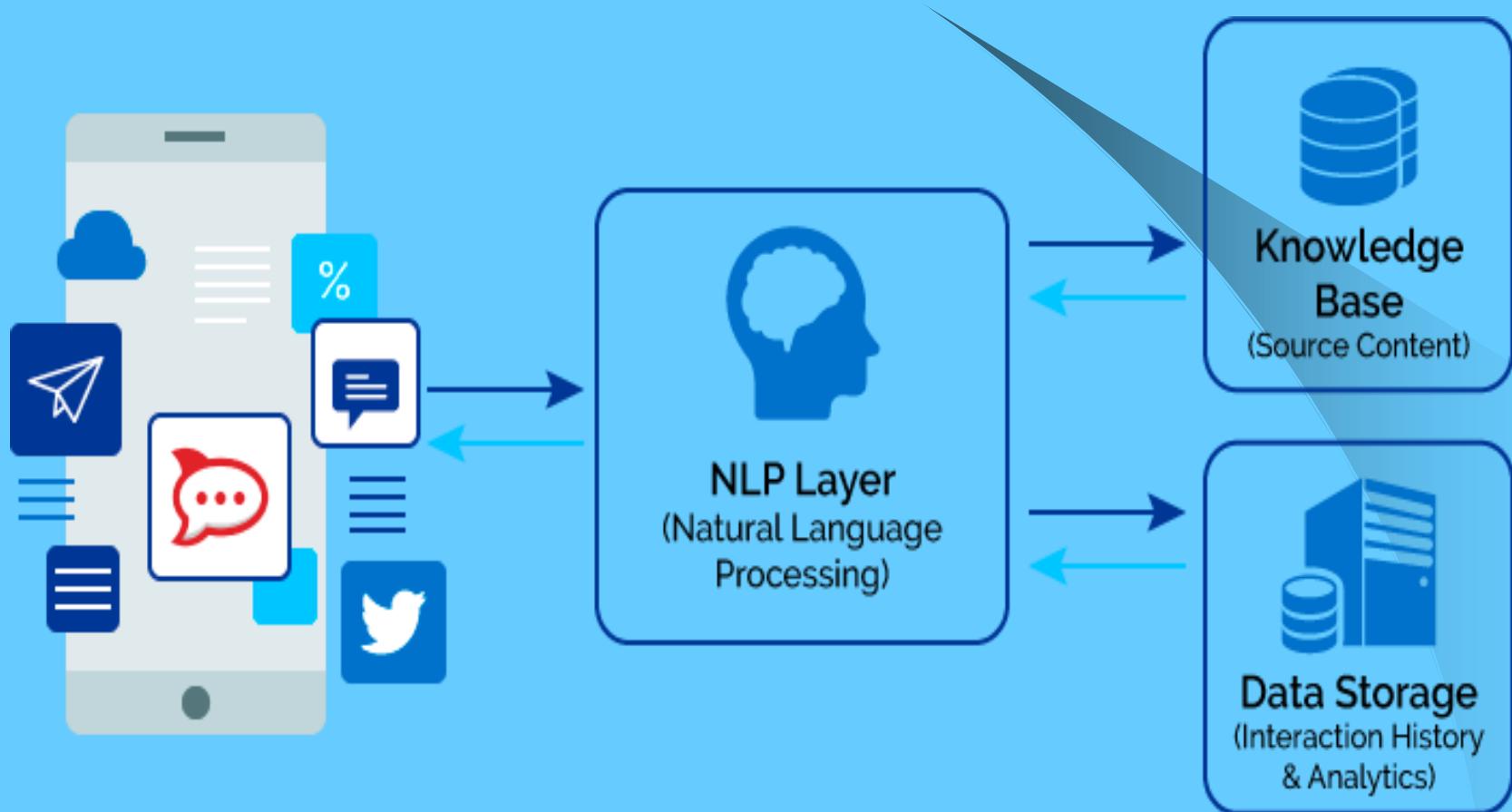
Found total 2,589 results.

Left	Key	Right
주말에 영화를 보러	갔다	왔어요.
저녁에 한국 영화를 보려	갈	거예요.
타나카 씨는 한국	가요	중에서 어떤 가수를 제일 좋아하니까?
그 행사에	가면	한국에서 인기 있는 드라마도 볼 수 있습니다.
집에	가면	제일 먼저 뭘 해요?
집에	가면	옷을 먼저 갈아입어요.
약을 먹거나 병원에	가요	.
저는 하와이에	가	봤어요.
학교에 올 때 입는 옷과 결혼식에	갈	때 입는 옷이 같을까요?
겨울 시즌	가	날씨가 우울하네요.

Vietnamese

	Key	Right
Tôi đã	đã	xem phim vào cuối tuần.
Tôi dự kiến sẽ	đi	xem phim Hàn Quốc vào buổi tối.
-	-	Trong nền âm nhạc Hàn Quốc thì bạn Tanaka thích ca sĩ
Nếu đi đến	đến	sự kiện đó thì có thể xem phim truyền hình đang được
Khi	đến	về đến nhà việc đầu tiên làm là gì?
Khi về đến	đến	nha thì thay quần áo trước.
Uống thuốc hoặc là	đã	đến bệnh viện.
Tôi đã từng	đã	đến Hawai.
Áo mặc khi	đã	đến trường và áo mặc khi đi đám cưới thì có khác nhau
nhé	đã	không?

# 10. Chatbot: ChatGPT et al.



# ChatGPT: Giải đề thi môn Văn THPT QG-2023

DU

Chỉ ra những từ ngữ, hình ảnh miêu tả cơn giông mùa hè trong những dòng thơ sau:

Tiếng sấm gõ trên bầu trời thật thấp

Gió từ đất thổi lên rất猛

Cát bay, lá bay, đá bay



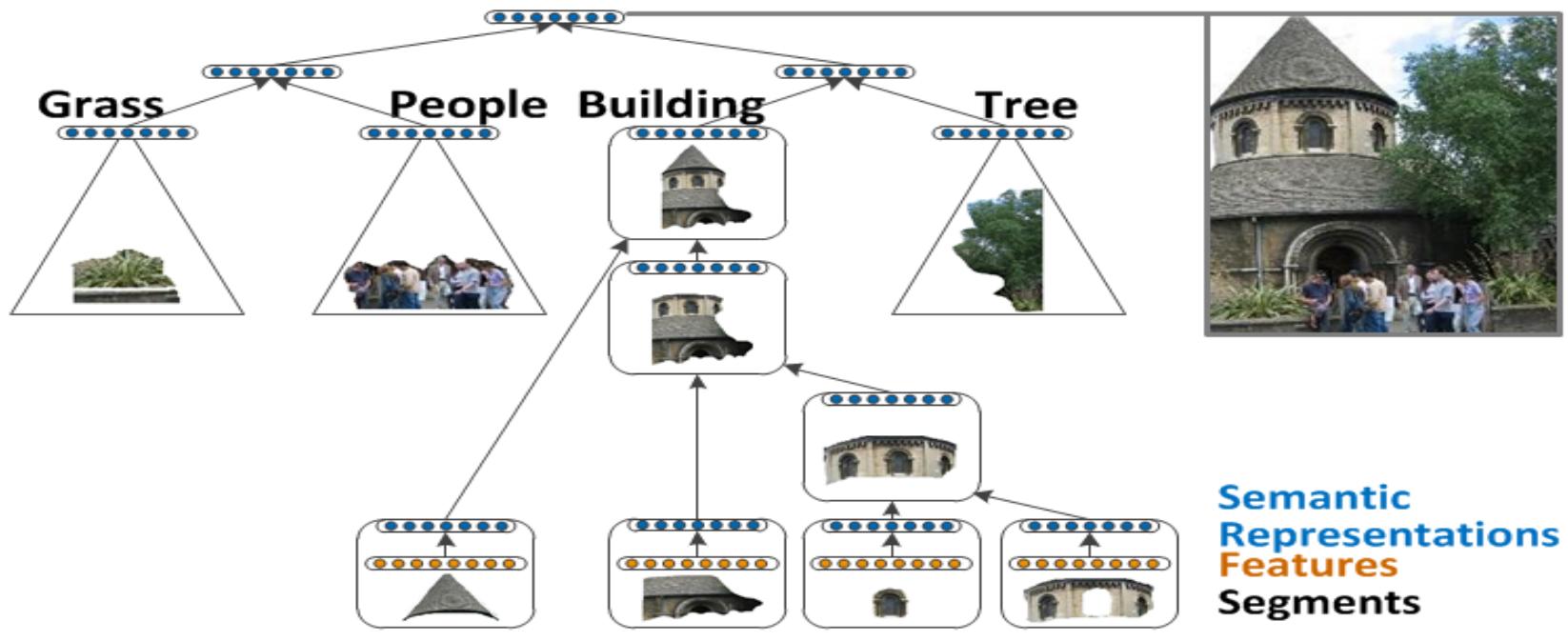
Trong đoạn thơ trích dưới đây, có những từ ngữ và hình ảnh miêu tả cơn giông mùa hè như sau:



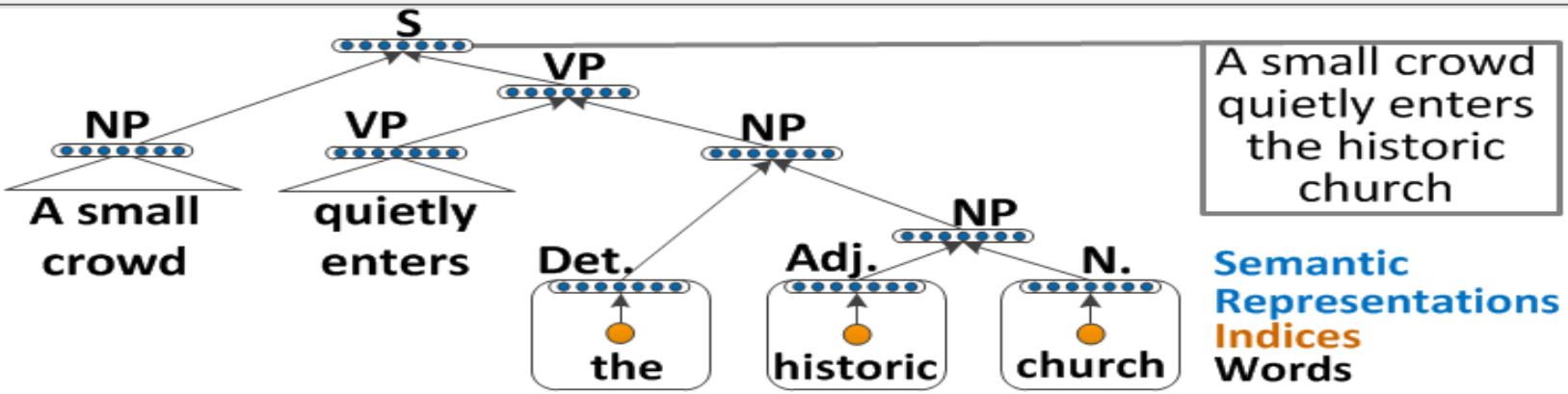
- **Tiếng sấm gõ trên bầu trời thật thấp:** Hình ảnh của tiếng sấm gõ trên bầu trời thấp tạo ra một âm thanh mạnh mẽ và xa xa, mang đến sự cảm giác của cơn giông đang đến gần.
- **Gió từ đất thổi lên rất猛:** Miêu tả về gió từ đất thổi lên và cảm nhận của người viết khi gió mạnh đến mức rất猛, cho thấy sức mạnh và cường độ của cơn giông.
- **Cát bay, lá bay, đá bay:** Hình ảnh của các vật liệu như cát, lá và đá được cuốn đi trong giông bão, tạo nên cảm giác hỗn loạn và mạnh mẽ của cơn giông mùa hè.

Những hình ảnh này tạo ra một bức tranh sống động về cơn giông mùa hè với âm thanh, sức mạnh và sự hỗn loạn của nó.

## Parsing Natural Scene Images



## Parsing Natural Language Sentences



# Computational Models:

---

**Artificial Intelligence:** A branch of computer science dealing with the simulation of intelligent behavior

**Machine Learning:** is a type of artificial intelligence ([AI](#)) that allows software applications to become more accurate at predicting outcomes via [training data](#).

- **Deep Learning:** requires big data
- Computational Linguistics: Data = Corpus
- **Corpus:** 语料库/yǔ liào kù/ “ngữ liệu khô”
- Corpus = Collection of spoken/written text
- Building Corpus: by [native-speaker](#), Master in [Applied Linguistics](#) (Computational Linguistics), Data Science.

# Corpus:

- PTB (Penn Tree Bank): [Pierre/NNP Vinken/NNP],/, [61/CD years/NNS] old/JJ ,/, will/MD join/VB [the/DT board/NN] as/IN [a/DT nonexecutive/JJ director/NN Nov./NNP 29/CD]./.
- CTB (Chinese Tree Bank): <S ID=12>( (IP-HLN (NP-SBJ (NN 外商) (NN 投资) (NN 企业)) (VP (VV 成为) (NP-OBJ (NP (NP-PN (NR 中国)) (NP (NN 外贸)))) (ADJP (JJ 重要)) (NP (NN 增长点)))) ) </S>
- (VTB: Vietnamese Tree Bank): <SEG id="1">  
Nguyên\_nhân/Nn/O là/Vc/O bão/Nn/O số/Nn/O 10/An/O  
đang/R/O chịu/Vv/O ảnh\_hưởng/Nn/O bởi/Cp/O  
hệ\_thống/Nn/O trực/Nn/O rãnh/Nn/O cao/Aa/O và/Cp/O  
sự/Nc/O lôi\_kéo/Vv/O từ/Cm/O siêu\_bão/Nn/TRM\_B  
Melor/Nr/TRM\_I\_ở/Cm/O ngoài/Cm/O khơi/Nn/O  
Philippines/Nr/LOC\_B ./PU/O</SEG

# Corpus:

[ Many/JJ styles/NNS ]

have/VBP

[ perforations/NNS ]

and/CC

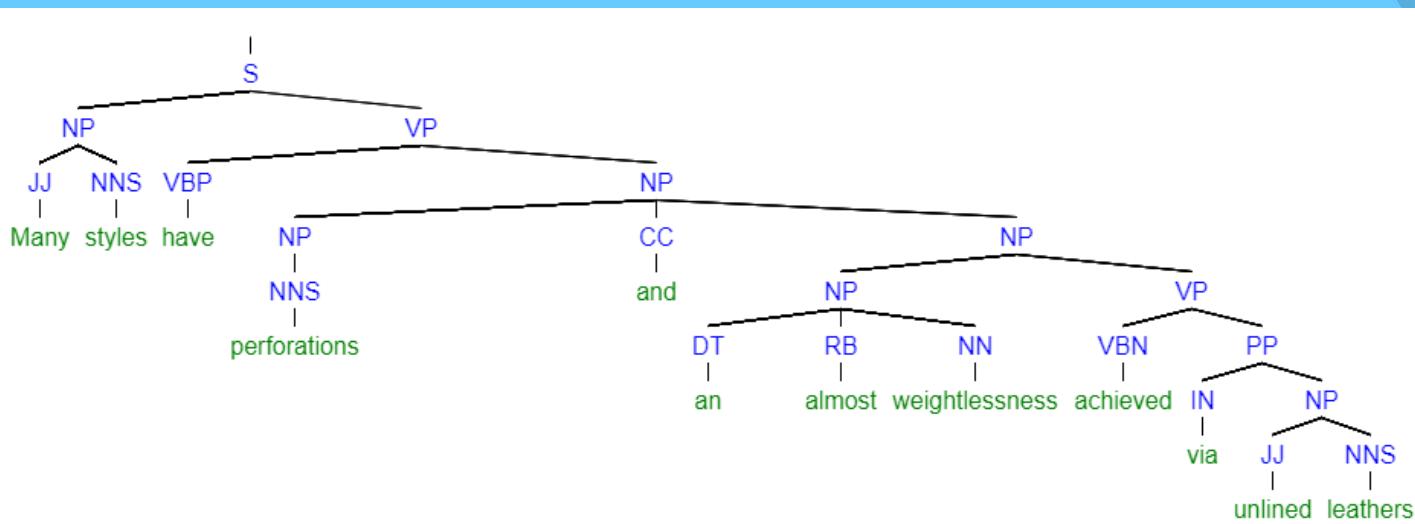
[ an/DT almost/RB weightlessness/NN ]

achieved/VBN via/IN

[ unlined/JJ leathers/NNS ]

./.

```
( (S - - -  
  (NP (JJ Many) (NNS styles) )  
  (VP (VBP have)  
   (NP  
    (NP (NNS perforations) )  
    (CC and)  
    (NP  
     (NP (DT an) (RB almost) (NN weightlessness) )  
     (VP (VBN achieved)  
      (PP (IN via)  
       (NP (JJ unlined) (NNS leathers) ))))))))
```



# Data production

THE NEW NEW WORLD

## *How Cheap Labor Drives China's A.I. Ambitions*



Workers at the headquarters of Ruijin Technology Company in Jiaxian, in central China's Henan Province. They identify objects in images to help artificial intelligence make sense of the world. Yan Cong for The New York Times

Data is the new oil, it has been said for years now. If data is the new oil, then China is already the largest producer with its factories packed with laborers working hard to annotate images and data for machine learning (*Analytics India Magazine*).

# Vietnamese Computational Linguistics

Rank	Language Name	Primary Country	Population
1	CHINESE, MANDARIN	China	885,000,000
2	SPANISH	Spain	332,000,000
3	ENGLISH	United Kingdom	322,000,000
4	BENGALI	Bangladesh	189,000,000
5	HINDI	India	182,000,000
6	PORTUGUESE	Portugal	170,000,000
7	RUSSIAN	Russia	170,000,000
8	JAPANESE	Japan	125,000,000
9	GERMAN, STANDARD	Germany	98,000,000
10	CHINESE, WU (Ngô)	China	77,175,000
11	JAVANESE	Indonesia, Java, Bali	75,500,800
12	KOREAN	Korea, South	75,000,000
13	FRENCH	France	72,000,000
14	VIETNAMESE	Vietnam	67,662,000
15	TELUGU	India	66,350,000
16	CHINESE, YUE (Việt)	China	66,000,000

# Abilities of Vietnamese language perception?

## Việt Nam giành huy chương Vàng môn kiếm chém đồng đội

VietNam+ 06/12/2019 38 liên quan

Đây là tấm huy chương Vàng thứ 4 của đoàn thể thao Việt Nam trong ngày thi đấu chính thức thứ 6 của SEA Games 30.

Thích

Bình luận

Chia sẻ

Á hậu, MC bình thản khai nhận  
nhiều lần bán dâm ngàn USD cho  
chách ở trụ sở công an

Viết Dũng - Theo Trí Thức Trẻ, 06/09/2018 17:34



**BÙI TIẾN DŨNG THỔ LỘ  
VIỆC VỢ CÓ BẦU VỚI CỰU  
HLV TRƯỞNG ĐỘI TUYỂN  
VIỆT NAM TẠI SÂN HÀNG  
ĐẤY**

[ttvn.vn](#) | 07/07/2019 12:00 AM

Bùi Tiến Dũng và Viettel đã có chiến thắng tối thiểu 1-0 trước CLB TPHCM ở vòng 14 V.League 2019 diễn ra vào tối 7/7. Sau trận đấu, anh cũng có cuộc gặp mặt ngắn với



# => Our strong point: Vietnamese-native speakers

Inquiry about Machine Translation for Vietnamese Inbox x   

임행선 <hs00.lim@samsung.com> 8/9/13 Star Up Down

to me

Dear Professor Dinh Dien,

This is the Software Center at Samsung Electronics in Korea. Our lab is currently researching the development of Korean <-> Vietnamese machine translation. While searching for Vietnamese universities and companies which have expertise in MT, we came across your name. We wonder whether you have conducted research on MT for Vietnamese language, and whether you have an ongoing research or project. If you share with us how things are with you, it will very helpful to us.

We also need the info on MT companies which work on Vietnamese. If you know any company or institution which supports Vietnamese with its own MT engine, please let us know.

Thank you in advance.

Best regards,  
Haengsun Eunice Lim

**Haengsun Eunice Lim**  
**Mobile. +82-10-2320-5040 / Tel. +82-31-279-2395**  
**E-mail. [hs00.lim@samsung.com](mailto:hs00.lim@samsung.com)**

**Web Platform Lab/ Software Center**  
**Samsung Electronics Co., LTD in Suwon, Korea**

## Inquiry about Machine Translation for Vietnamese

Inbox x



임행선 <hs00.lim@samsung.com>

to me ▾

8/9/13



Dear Professor Dinh Dien,

This is the Software Center at Samsung Electronics in Korea. Our lab is currently researching the development of Korean <-> Vietnamese machine translation.

While searching for Vietnamese universities and companies which have expertise in MT, we came across your name.

We wonder whether you have conducted research on MT for Vietnamese language, and whether you have an ongoing research or project. If you share with us how things are with you, it will be very helpful to us.

We also need the info on MT companies which work on Vietnamese. If you know any company or institution which supports Vietnamese with its own MT engine, please let us know.

Thank you in advance.

Best regards,

Haengsun Eunice Lim

**Haengsun Eunice Lim**

**Mobile. +82-10-2320-5040 / Tel. +82-31-279-2395**

**E-mail. [hs00.lim@samsung.com](mailto:hs00.lim@samsung.com)**

**Web Platform Lab/ Software Center**

**Samsung Electronics Co., LTD in Suwon, Korea**

**Sent:** Tuesday, January 19, 2016 1:58 PM

**To:** [ddien@fit.hcmus.edu.vn](mailto:ddien@fit.hcmus.edu.vn)

**Subject:** Acquiring Vietnamese treebank

Dear Prof. Dinh Dien,

HyunJeong Choe <[hyunjeongc@google.com](mailto:hyunjeongc@google.com)>

1/21/16



to me

Thank you so much your prompt reply!

If your treebank contain 300k, then we would like to acquire the entire set.

We are Natural language understanding team under Google research team and focusing on several NLP projects. We'd like to use your treebank to train our Vietnamese segmenter, PoS tagger and NER tagger. These analyzer will be used several Google projects such as conversational search.

The Licensee may use the data internally only. The Licensee may not:

1. Distribute the data;
2. Publish any research in which the data was used without providing a citation acknowledging that the data was developed by the Computation Linguistics Center of HCMUS.

Best,

-HJ

**Date:** 15-Oct-2015

**From:** Kohei Saito <AdvancedLinguistics@gmail.com>

**Subject:** Vietnamese; Computational Linguistics; Morphology; Phonology; Semantics; Syntax: Analytic Linguistic Project Manager, Google, Inc., Singapore

University or Organization: **Google, Inc.**

Department: Natural Language Understanding

Job Location: Singapore, Singapore

Job Title: Analytic Linguistic Project Manager [Vietnamese]

Job Rank: Analytic Linguistic Project Manager; Manager

Specialty Areas: Computational Linguistics; Morphology; Phonology; Semantics; Syntax

Required Language(s): **Vietnamese (vie)**

Description:

The role of the Analytic Linguistic Project Manager is to consult with Natural Language Understanding Researchers on creating guidelines and setting standards for a variety of NLP projects as well as to manage the work of a team of junior linguists to achieve high quality data output.

This includes:

- Training, managing and overseeing the work of a team of junior linguists
- Creating guidelines for semantic, syntactic and morphological projects
- Evaluating and analyzing data quality
- Consulting with researchers and engineers on the development of linguistic databases

Job requirements:

- **Native-level speaker of Vietnamese** and fluent in English
- **Master's degree or higher in Linguistics or Computational Linguistics**, specializing in semantics, syntax, morphology or lexicography
- Ability to quickly grasp technical concepts; should have an interest in natural language processing
- Excellent oral and written communication skills
- Good organizational skills
- Previous project management and people management experience preferred
- Some programming language or previous experience working in a Linux environment a plus

Hivan Fagnano hivan.fagnano

Số hóa > Công nghệ

Thứ tư, 30/9/2020, 13:00 (GMT+7)

to clc 

Greetings,

Our company is looking for  
well known multinational co  
We're building up a team of  
studies.

The project, which would last 3-5 months, involves performing quality control tasks of audio-recorded files \ voice overs in the linguists' native language, so phonetic transcription, pronunciation transcription and proofreading skills are required.

Please note that the right candidates must be native speakers of Vietnamese.

By visiting your site, I've noticed the the Computational Linguistic Center focuses on spelling checker, grammar checker, Text Translation, Contrastive Linguistics, etc. and I thought students from your course might be considered good candidates, as we need native Vietnamese speakers with the above mentioned skills.

Would it be possible talking with teachers from the Center?

## Apple tuyển người nói tiếng Việt làm Siri

Apple đăng tuyển nhân sự thành thạo tiếng Việt trên trang tuyển dụng của mình, nhiều khả năng sẽ phát triển trợ lý ảo Siri cho thị trường Việt Nam.

Trên trang tuyển dụng, Apple mới bổ sung vị trí chuyên viên Ngôn ngữ Việt Nam cho mảng trí tuệ nhân tạo và học máy. Người được tuyển dụng sẽ làm việc trong đội ngũ phát triển Siri, tại văn phòng ở khu Ang Mo Kio (Singapore). Siri là trợ lý ảo của Apple và là một trong những ứng dụng thực tế nhất về AI mà Apple đang phát triển.

# MY PRODUCTS

## Dictionary

An entry in the Chinese-Vietnamese Dictionary:

```
<WORD>
  <HEAD>油然</HEAD>
  <PHONETIC>yóurán</PHONETIC>
  <BODY>
    <TXT_V>Tự nhiên</TXT_V>
    <EXAMPLE>
      <TXT_C>敬慕之心，油然而生</TXT_C>
      <TXT_V>Lòng ngưỡng mộ, tự nhiên mà có</TXT_V>
    </EXAMPLE>
  </BODY>
  <BODY>
    <TXT_V> hơi nước bốc lên</TXT_V>
    <EXAMPLE>
      <TXT_C>油然作云</TXT_C>
      <TXT_V>Hơi nước bốc lên thành mây</TXT_V>
    </EXAMPLE>
  </BODY>
</WORD>
```

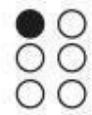
An image of the Oxford Advanced Learner's Dictionary, showing its blue and white cover with the title and a small image of the British Parliament building.

An advertisement for the Kim Từ Điển electronic dictionary. It features a small handheld device with a screen and keyboard, labeled 'GD7200M'. Text on the ad includes 'Kim Từ Điển', 'Dịch Câu & Phát Âm Thông Minh', and 'ANH - VIỆT - PHÁP - HOA - NHẬT - HÀN - ĐỨC - NGA'.

A close-up image of the Kim Từ Điển GD7200M electronic dictionary, showing its open screen displaying various icons and text, and its physical keyboard and body.



# KÝ HIỆU CHỮ BRAILLE VIỆT NGỮ



A



Ă



Â



B



C



Đ



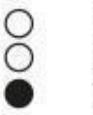
D



E



Ê



F



G



H



I



J



K



L



M



N



O



Ô



Õ



P



Q



R



S



T



U



Ú



Û



Ý



ÿ



ñ



á



é



ô



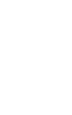
õ



sắc



huyền



hỏi



ngã



năng



hai chấm :



phẩy ,



chấm phẩy ;



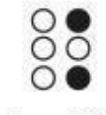
chấm câu .



chấm thang !



chấm hỏi ?



Báo viết hoa



(



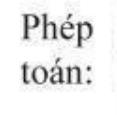
)



mở "



đóng "



Phép  
tính:



+



-



X



:



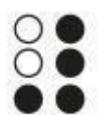
=



<



>



Báo số



0



1



2



3



4



5



6



7



8



9

Trung tâm dữ liệu đa ngữ  
Kim Từ Điện (KMDC) chúng  
tôi chuyên sản xuất các  
Phần mềm có hỗ trợ tiếng  
Việt cho người khiếm thị.





# Language Technology

making good progress

mostly solved

## Spam detection

Let's go to Agra!  
Buy V1AGRA ...



## Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

## Named entity recognition (NER)

PERSON	ORG	LOC
Einstein	met with	UN officials in Princeton

## Sentiment analysis

Best roast chicken in San Francisco!  
The waiter ignored us for 20 minutes.



## Coreference resolution

Carter told Mubarak he shouldn't run again.

## Word sense disambiguation (WSD)

I need new batteries for my **mouse**.



## Parsing

I can see Alcatraz from the window!

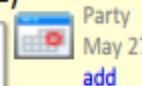
## Machine translation (MT)

第13届上海国际电影节开幕...  
The 13<sup>th</sup> Shanghai International Film Festival...



## Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30



still really hard

## Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

## Paraphrase

XYZ acquired ABC yesterday  
ABC has been taken over by XYZ

## Summarization

The Dow Jones is up  
The S&P500 jumped  
Housing prices rose



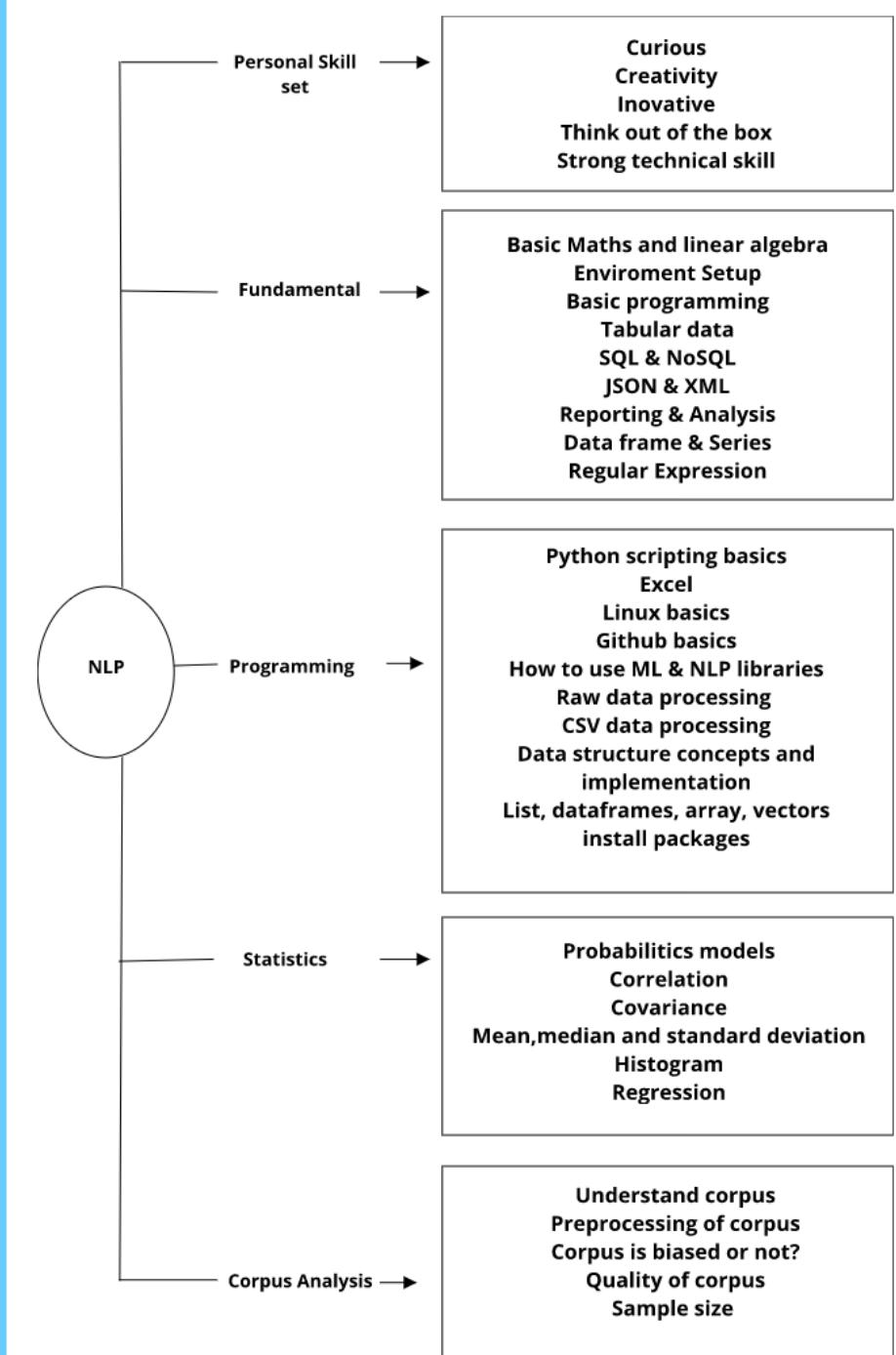
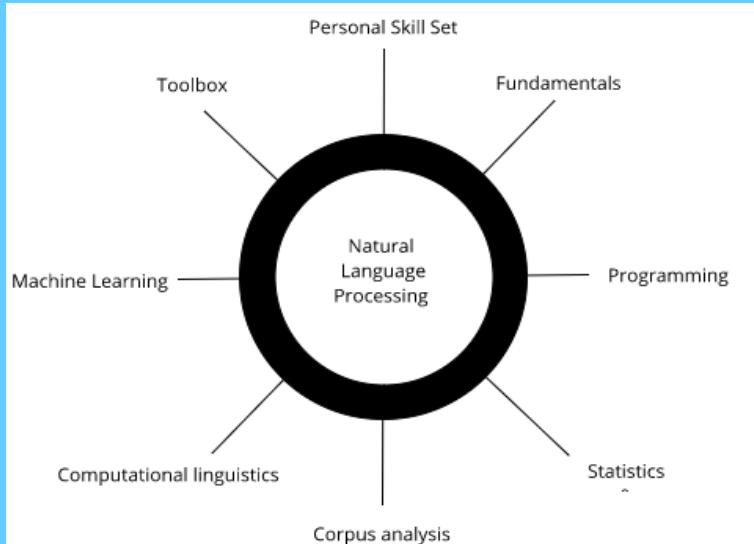
Economy is good

## Dialog

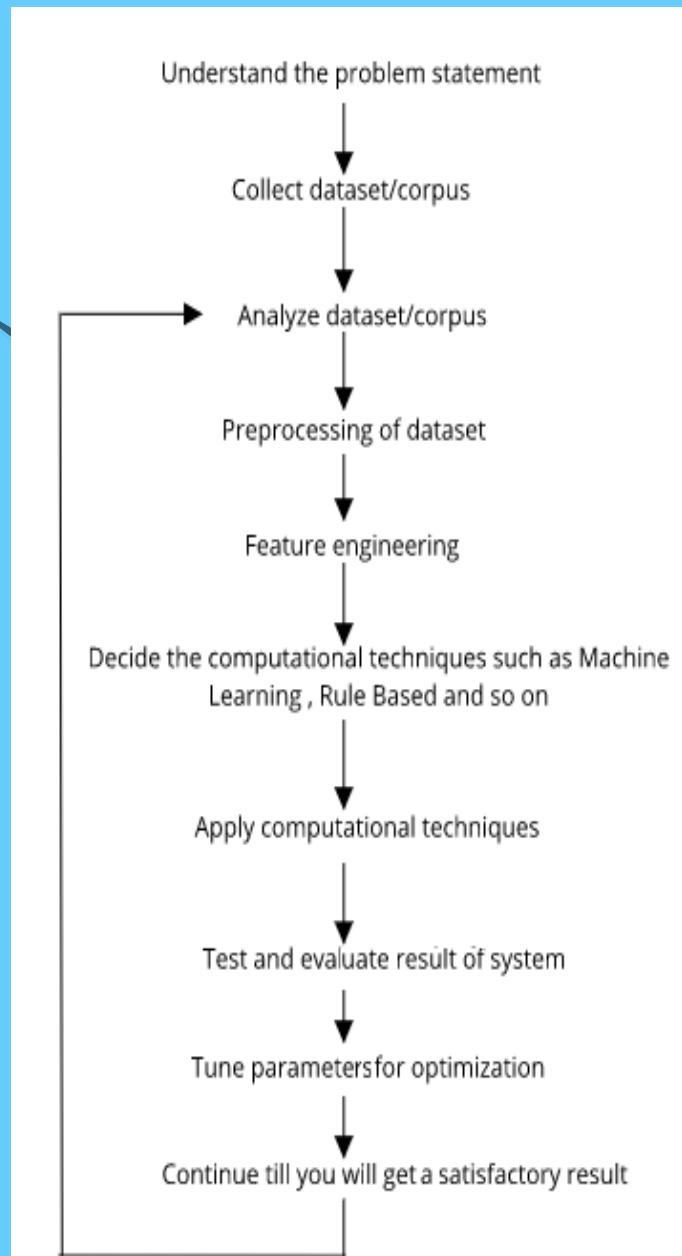
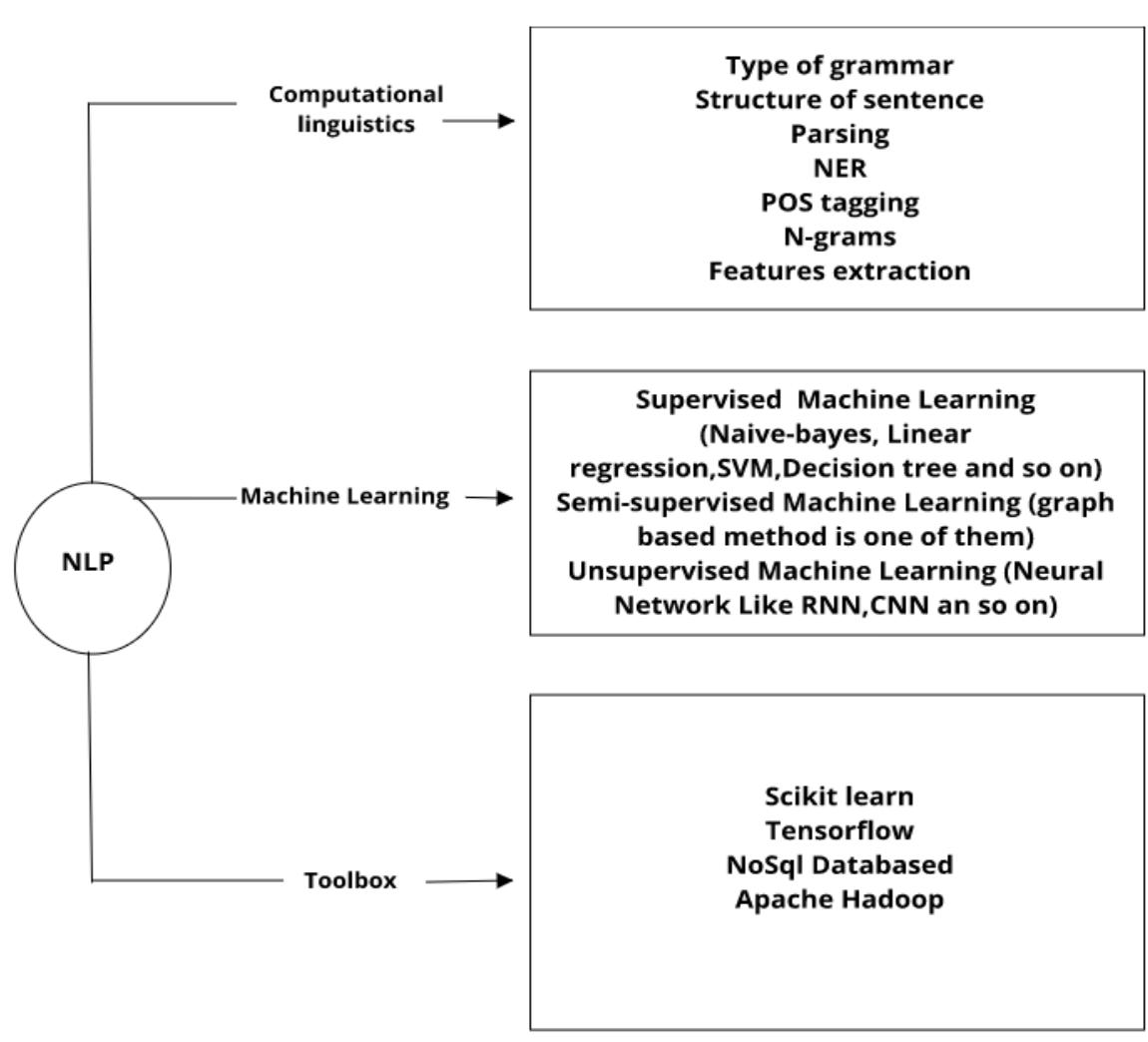
Where is Citizen Kane playing in SF?  
Castro Theatre at 7:30. Do you want a ticket?



# What do you need for NLP?



# What will you learn in this NLP course?



# Course content

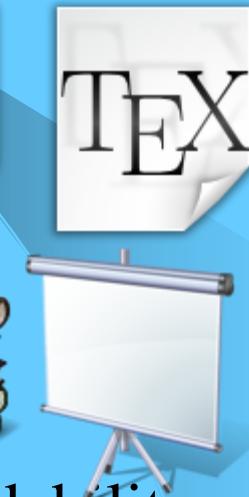
Introduction to Natural Languages

Review Formal Languages

Rule-based NLP: parsing

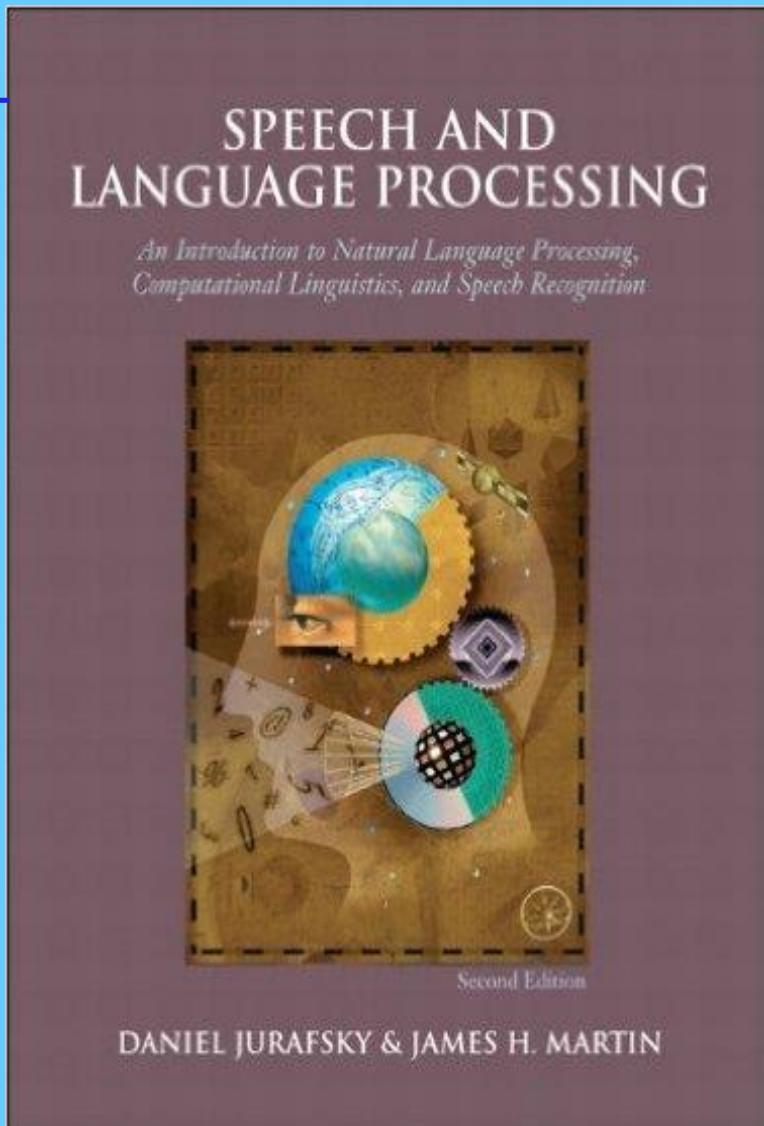
Corpus-based NLP: tagging

NLP Applications: text classification, readability, stylometry, mining, summarization, understanding, translation

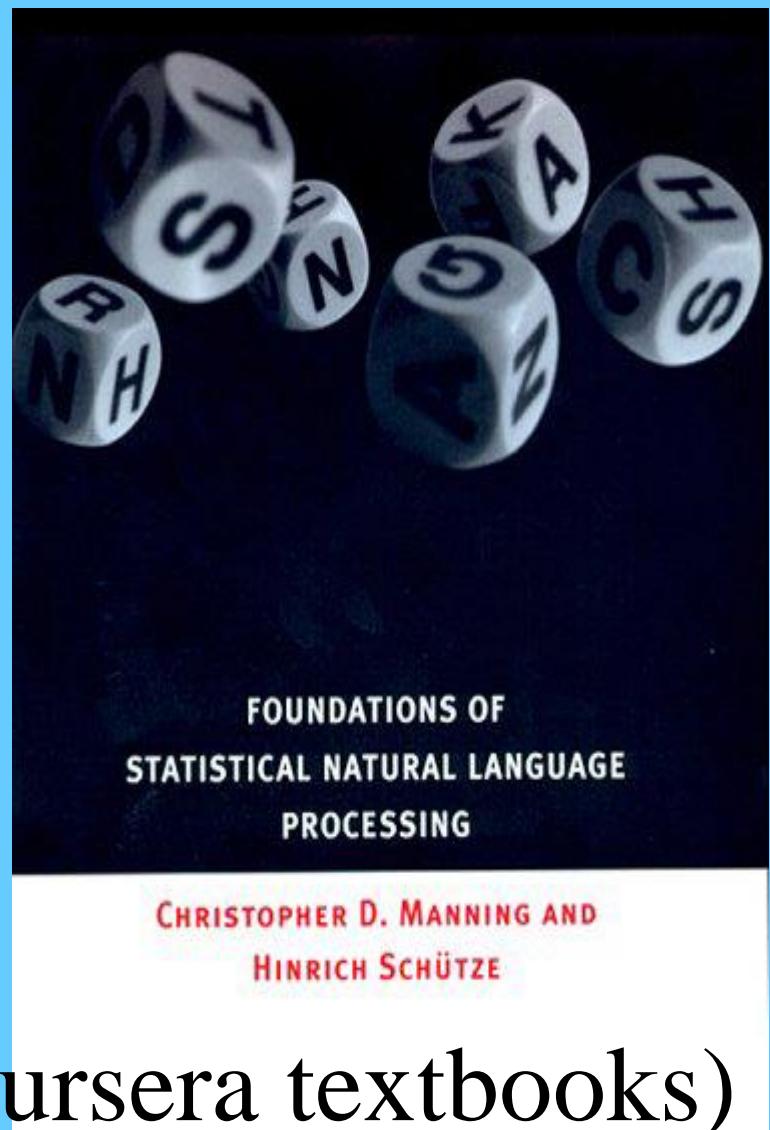


# Textbooks (theory)

01



02



(Coursera textbooks)

# Textbooks (practice)

03 **Python**

## Natural Language Processing

Explore NLP with machine learning  
and deep learning techniques

Jalaj Thanaki

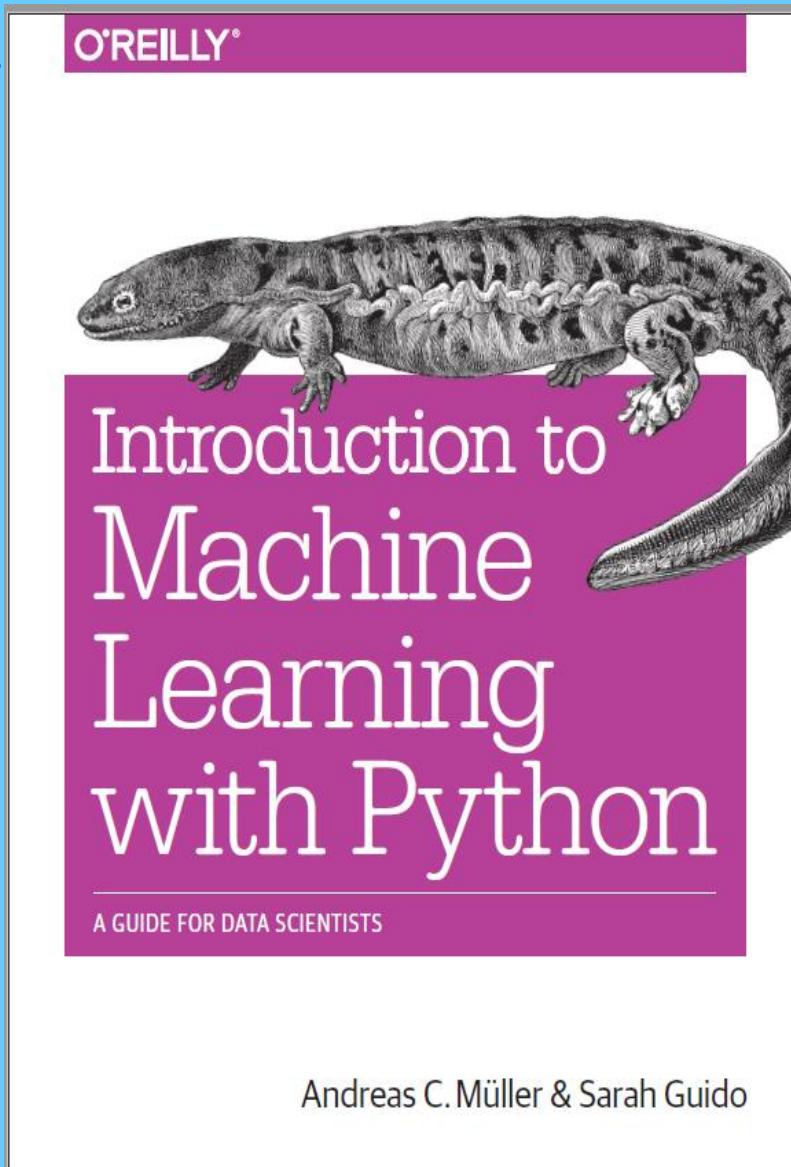
Packt

BIRMINGHAM - MUMBAI

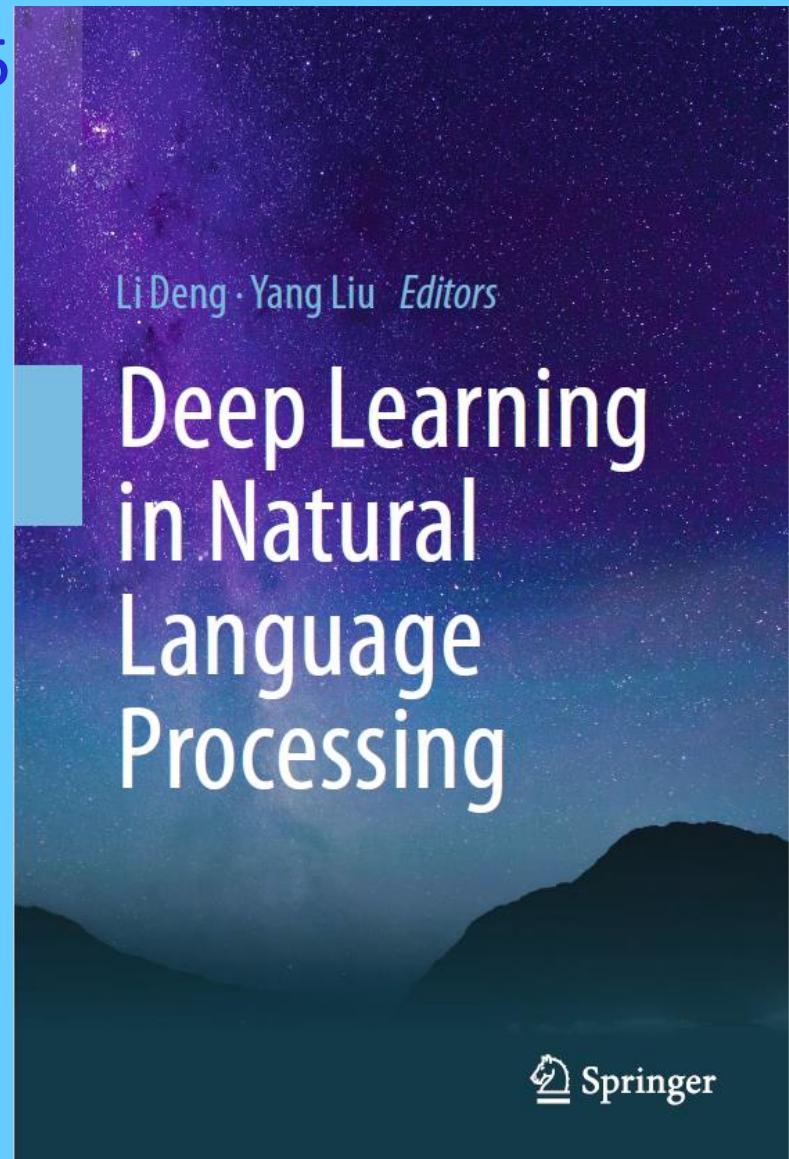
- >  Introduction
- >  Corpus & Dataset
- >  Structure of Sentences
- >  Preprocessing
- >  Feature Engineering & NLP Algorithms
- >  Advanced Feature Engineering & NLP Algorithms
- >  Rule-based System for NLP
- >  Machine Learning for NLP Problems
- >  Deep Learning for NLU & NLG Problems
- >  Advanced Tools
- >  Improve your NLP Skills
- >  Installation Guide

# Reference books (ML techniques)

04



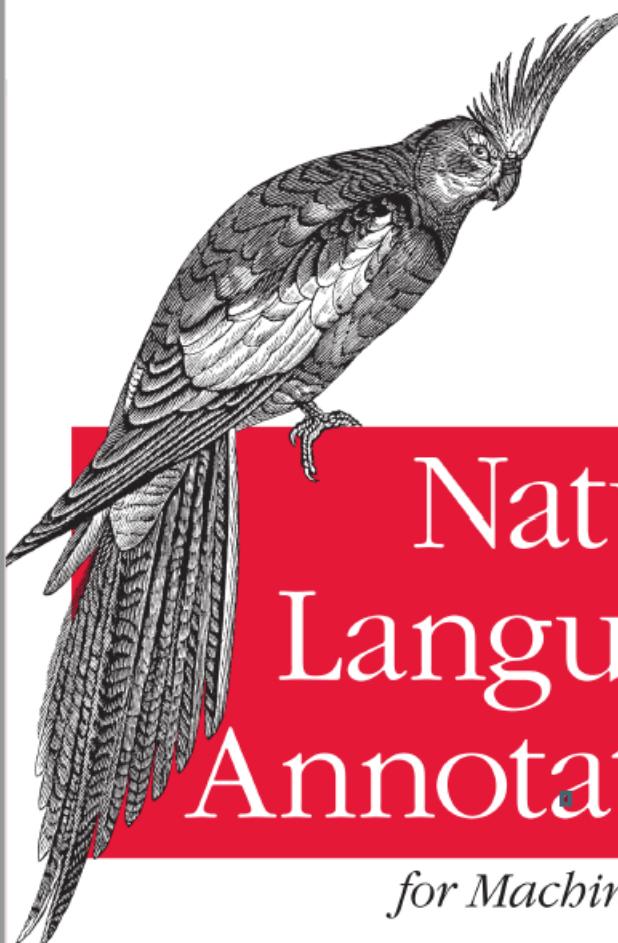
05



# Reference books (training corpus)

## 06

*A Guide to Corpus-Building for Applications*



- > Preface
- > Chapter 1. The Basics
- > Chapter 2. Defining Your Goal and Dataset
- > Chapter 3. Corpus Analytics
- > Chapter 4. Building Your Model and Specification
- > Chapter 5. Applying and Adopting Annotation Standards
- > Chapter 6. Annotation and Adjudication
- > Chapter 7. Training: Machine Learning
- > Chapter 8. Testing and Evaluation
- > Chapter 9. Revising and Reporting
- > Chapter 10. Annotation: TimeML
- > Chapter 11. Automatic Annotation: Generating TimeML
- > Chapter 12. Afterword: The Future of Annotation
- > Appendix A. List of Available Corpora and Specifications

# Reference books (linguistics)



MORGAN & CLAYPOOL PUBLISHERS

## 07 Linguistic Fundamentals for Natural Language Processing

*100 Essentials from  
Morphology and Syntax*

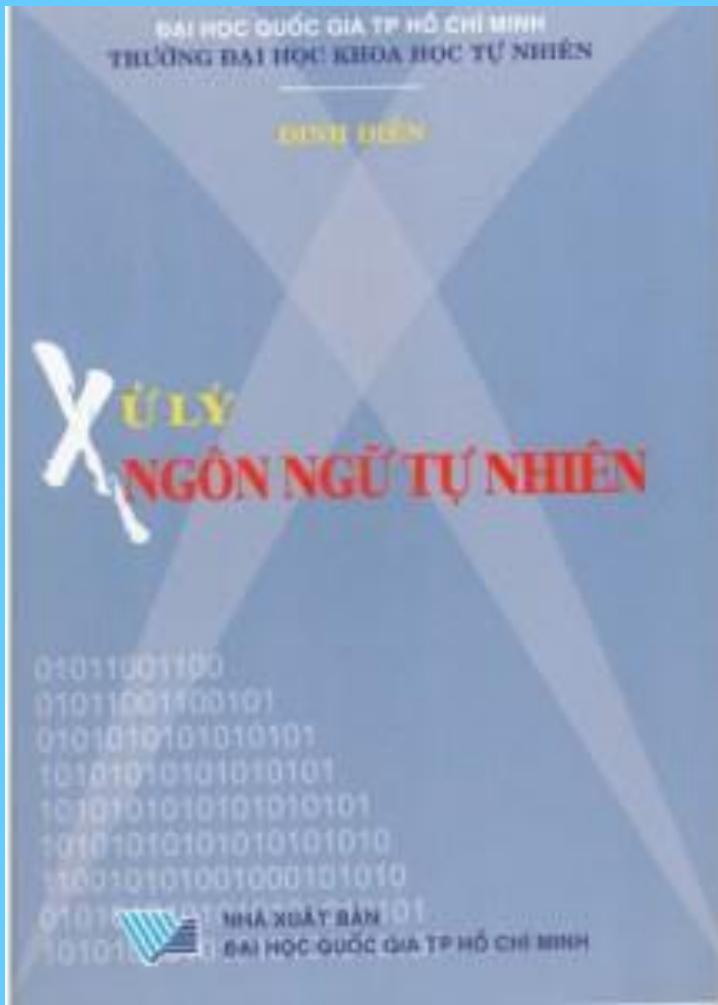
Emily M. Bender

**SYNTHESIS LECTURES ON  
HUMAN LANGUAGE TECHNOLOGIES**

Graeme Hirst, *Series Editor*

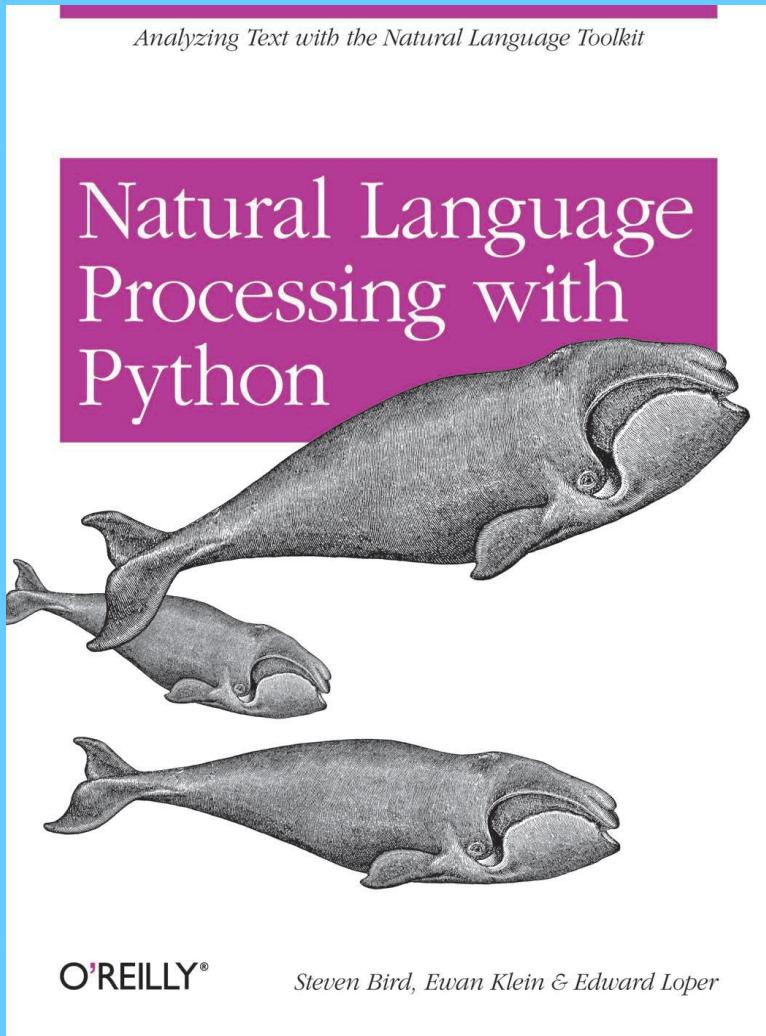
- > [Introduction/motivation](#)
- > [Morphology: Introduction](#)
- > [Morphophonology](#)
- > [Morphosyntax](#)
- > [Syntax: Introduction](#)
- > [Parts of speech](#)
- > [Heads, arguments and adjuncts](#)
- > [Argument types and grammatical functions](#)
- > [Mismatches between syntactic position and semantic roles](#)
- > [Resources](#)

# Reference books (Vietnamese)

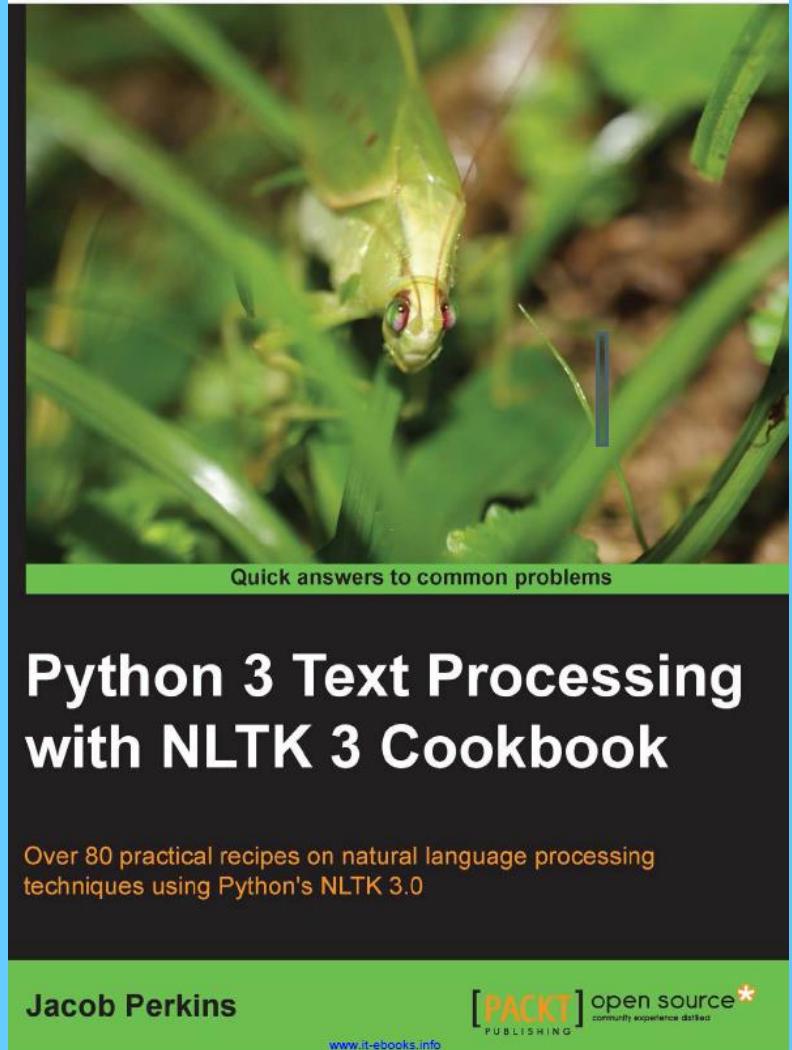


# Reference books (programming)

08



09



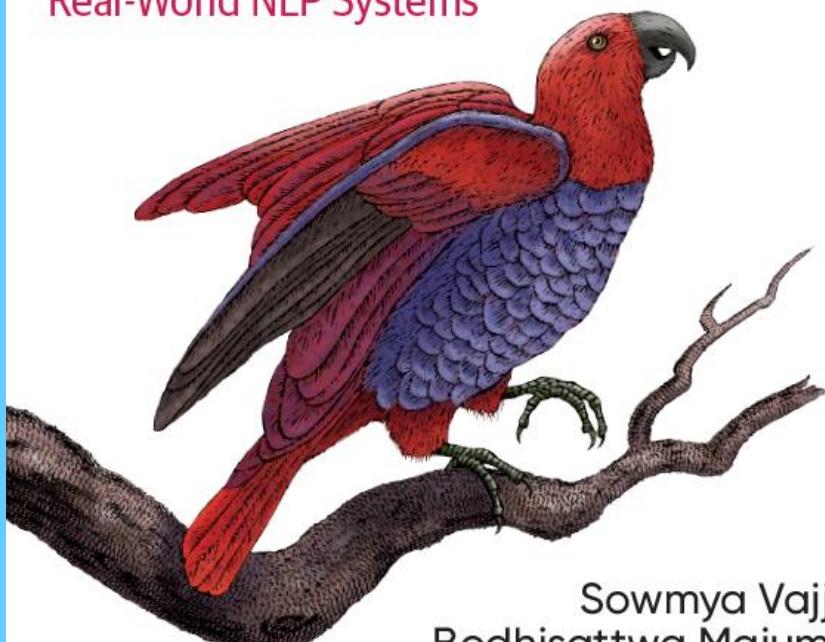
# Reference books (applications)

10

O'REILLY®

## Practical Natural Language Processing

A Comprehensive Guide to Building  
Real-World NLP Systems



Sowmya Vajjala,  
Bodhisattwa Majumder,  
Anuj Gupta & Harshit Surana

- > Preface
- ▽ Part I. Foundations
  - > Chapter 1. NLP: A Primer
  - > Chapter 2. NLP Pipeline
  - > Chapter 3. Text Representation
- ▽ Part II. Essentials
  - > Chapter 4. Text Classification
  - > Chapter 5. Information Extraction
  - > Chapter 6. Chatbots
  - > Chapter 7. Topics in Brief
- ▽ Part III. Applied
  - > Chapter 8. Social Media
  - > Chapter 9. E-Commerce and Retail
  - > Chapter 10. Healthcare, Finance, and Law
- ▽ Part IV. Bringing It All Together
  - > Chapter 11. The End-to-End NLP Process

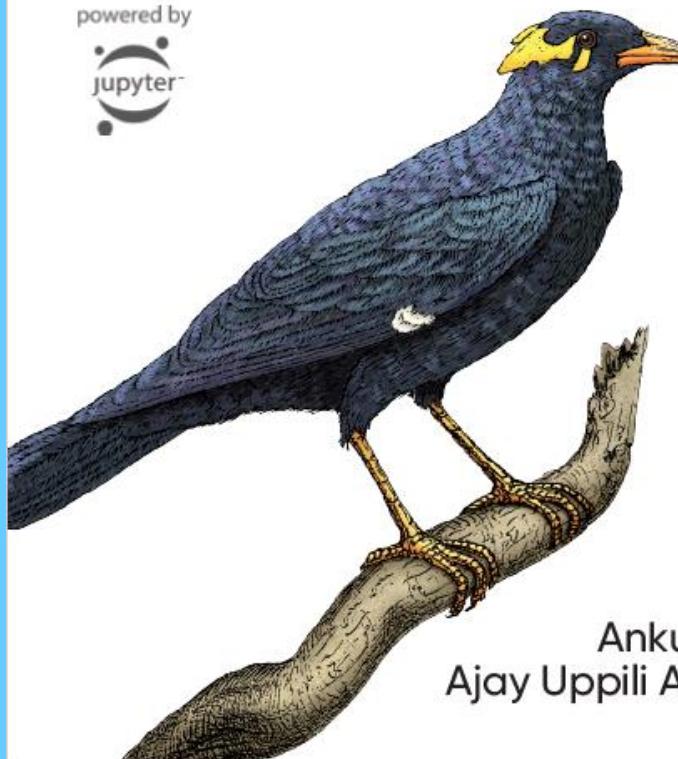
# Reference books (applications)

11 O'REILLY®

## Applied Natural Language Processing in the Enterprise

Teaching Machines to Read, Write & Understand

powered by



Ankur A. Patel &  
Ajay Uppili Arasanipalai

- ▼ Part I. Scratching the Surface
  - > Chapter 1. Introduction to NLP
  - > Chapter 2. Transformers and Transfer Learning
  - > Chapter 3. NLP Tasks and Applications
- ▼ Part II. The Cogs in the Machine
  - > Chapter 4. Tokenization
  - > Chapter 5. Embeddings: How Machines "Understand" Words
  - > Chapter 6. Recurrent Neural Networks and Other Sequence Models
  - > Chapter 7. Transformers
  - > Chapter 8. BERTology: Putting It All Together
- ▼ Part III. Outside the Wall
  - > Chapter 9. Tools of the Trade
  - > Chapter 10. Visualization
  - > Chapter 11. Productionization
  - > Chapter 12. Conclusion
- > Appendix A. Scaling

# Reference course (Coursera)

1. Natural Language Processing, Dan Jurafsky (Stanford University)

sentences shows context  
**probability**  
and grammar at times  
word model  
question contexts discourse  
can parsing based just three  
as is information  
STATS  
problem rule

## Introduction to NLP

What is Natural Language Processing?

MORE VIDEOS

→ Main reference slides: Stanford Uni.

# Reference websites (papers, dataset)

## Association for Computational Linguistics (ACL)

### **ACL Anthology** <https://aclanthology.org>

- A Digital Archive of Top Research Papers in CL, NLP.

<EACL> European Chapter of the ACL

<NAACL> North American chapter of the ACL

<EMNLP> Empirical Methods in Natural Language Processing

CoNLL: Conference on Computational Natural Language Learning

COLING: International Conference on Computational Linguistics

#### ❖ Vietnamese NLP:

<https://vlsp.org.vn/> (Vietnamese Language Speech Processing)

<https://www.clc.hcmus.edu.vn/> (Computational Linguistics

Center, Uni. of Science, HCMC-VNU).





# Computational Linguistics Journal (2022)

## Contents

- Computational Linguistics, Volume 48, Issue 1 - March 2022 [10 papers](#)
- Computational Linguistics, Volume 48, Issue 2 - June 2022 [8 papers](#)
- Computational Linguistics, Volume 48, Issue 3 - September 2022 [7 papers](#)

Show all abstracts ▾

↑up

### [bib \(full\)](#) Computational Linguistics, Volume 48, Issue 1 - March 2022

[pdf](#) [bib](#)

**Obituary: Martin Kay**

Ronald M. Kaplan | Hans Uszkoreit

[pdf](#) [bib](#) [abs](#)

**To Augment or Not to Augment? A Comparative Study on Text Augmentation Techniques for Low-Resource NLP**

Gözde Güл Şahin

[pdf](#) [bib](#) [abs](#)

**Novelty Detection: A Perspective from Natural Language Processing**

Tirthankar Ghosal | Tanik Saikh | Tameesh Biswas | Asif Ekbal | Pushpak Bhattacharyya

[pdf](#) [bib](#) [abs](#)

**Improved N-Best Extraction with an Evaluation on Language Data**

Johanna Björklund | Frank Drewes | Anna Jonsson

[pdf](#) [bib](#) [abs](#)

**Linguistic Parameters of Spontaneous Speech for Identifying Mild Cognitive Impairment and Alzheimer Disease**

Veronika Vincze | Martina Katalin Szabó | Ildikó Hoffmann | László Tóth | Magdolna Pákáski | János Kálman | Gábor Gosztolya

[pdf](#) [bib](#) [abs](#)

**Deep Learning for Text Style Transfer: A Survey**

Di Jin | Zhijing Jin | Zhiting Hu | Olga Vechtomova | Rada Mihalcea