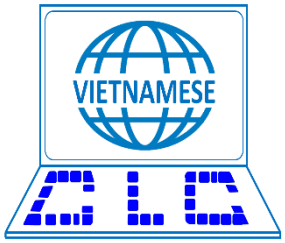
University of Science

Computational Linguistics Center

Natural Language Processing

Section 0:

Introduction to Natural Languages Processing



Lecturer: Assoc.Prof. Dr. Dinh Dien

ddien@fit.hcmus.edu.vn

Lecturer

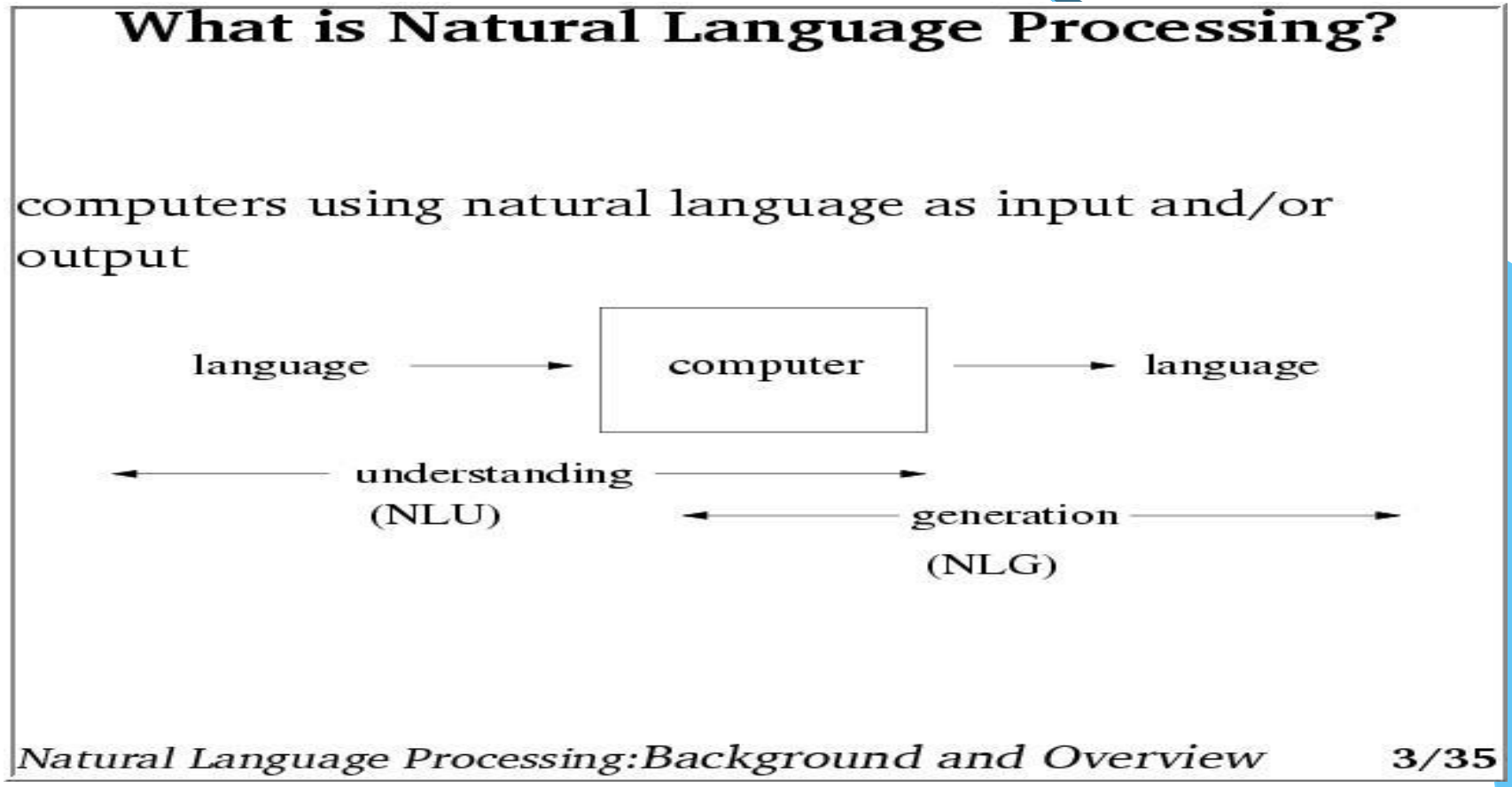
Teaching Assistant: Buu Long – An Vinh – Thuy Hang

Natural Language Processing (NLP)

- What is NLP ?
- Why NLP is hard?
- NLP applications
- What will this course be about?
- Textbooks
- Website - dataset
- Grading

What is NLP?

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



Why NLP is hard?

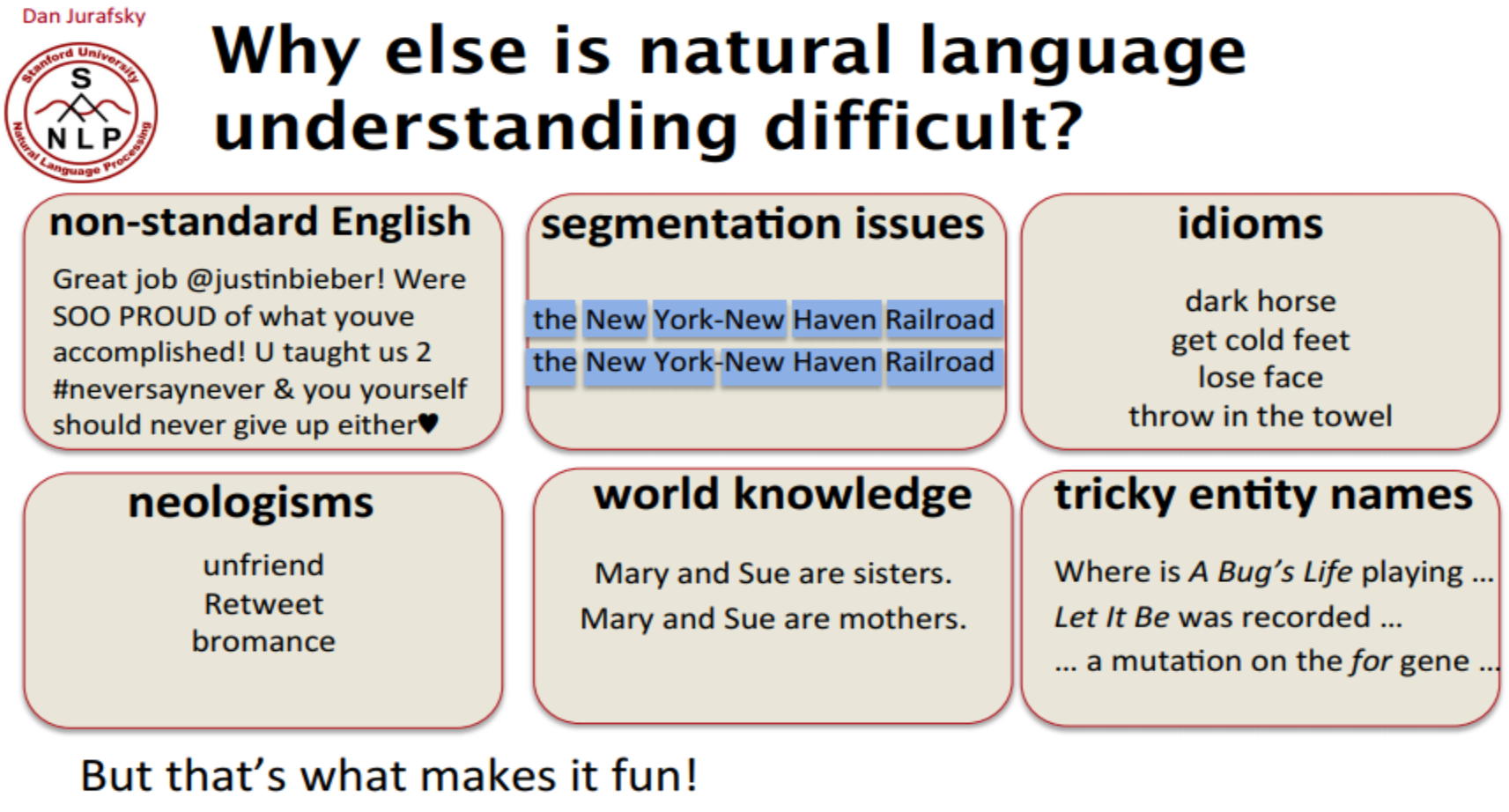
- Reason (1) – human language is ambiguous:
- Ex1 (pronoun resolution):
- Jack drank the wine on the table. ***It*** was red and round.
- Jack saw Sam at the party. ***He*** went back to the bar to get another drink.
- Jack saw Sam at the party. ***He*** clearly had drunk too much.
- Ex2: PrePosition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

Why NLP is hard?

- Reason (3) –Non-standard text, Idioms / Neologisms:

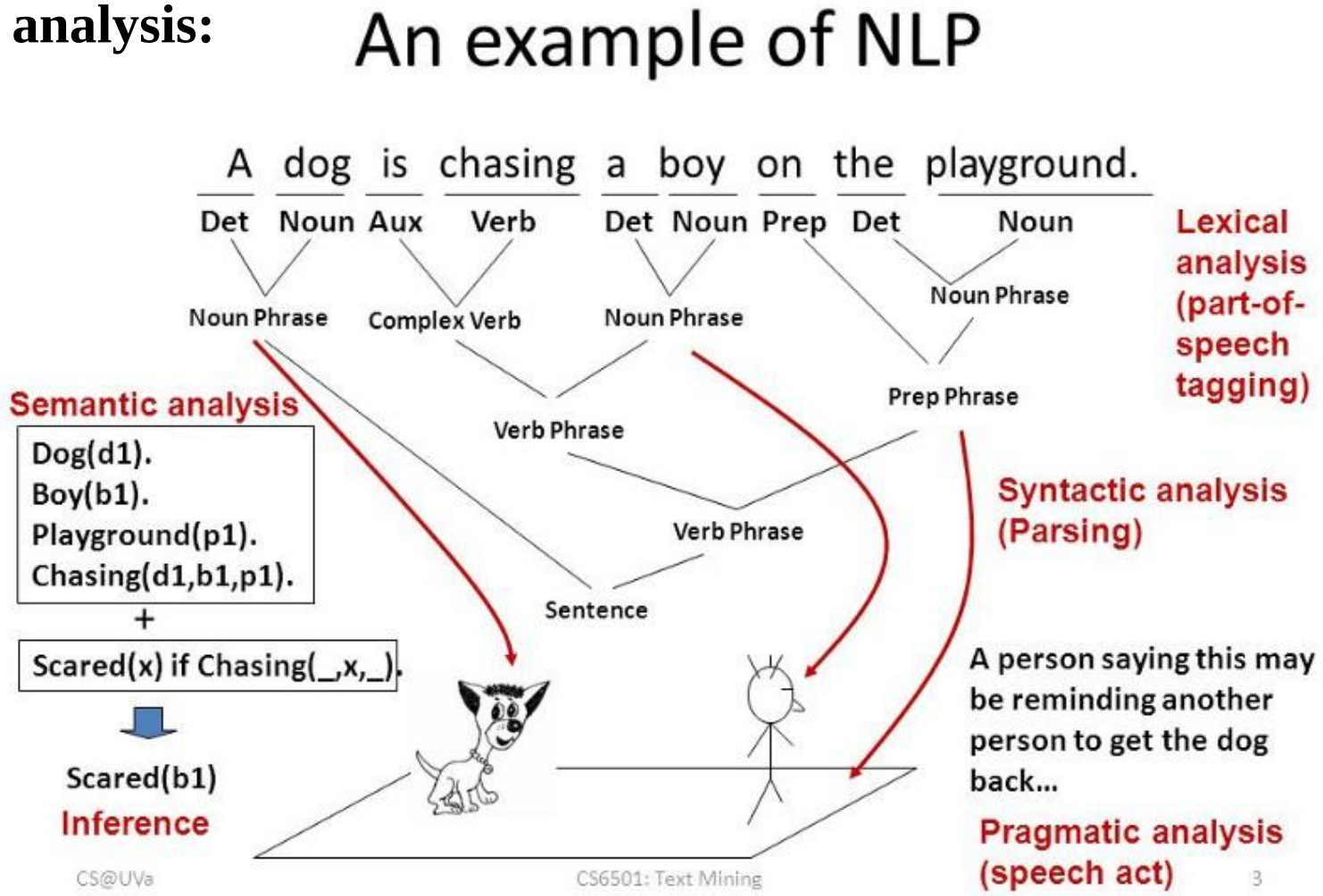


Why NLP is hard?

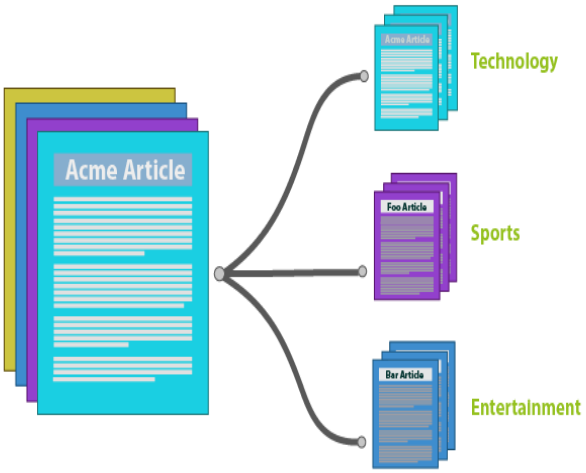
- Reason (4) – Language is difficult even for human.
- Learning mother tongue (native language)
  -- you might think it's easy, but…compare 5 year old vs. 10 year old vs. 20 year old
- Learning foreign languages
  – even harder

NLP APPLICATIONS

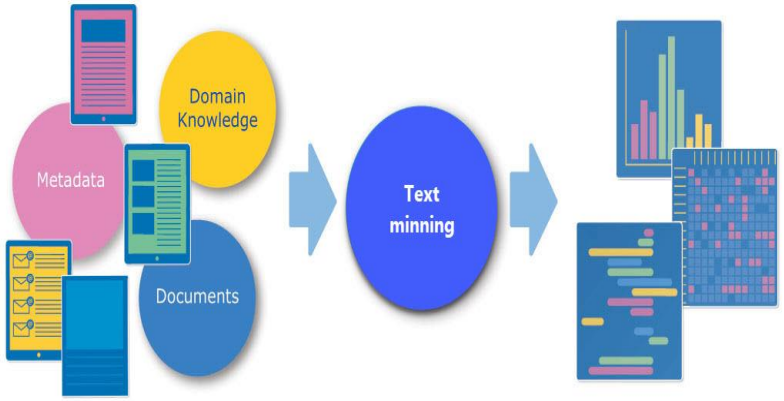
1. Linguistics analysis:



2. Text classification:

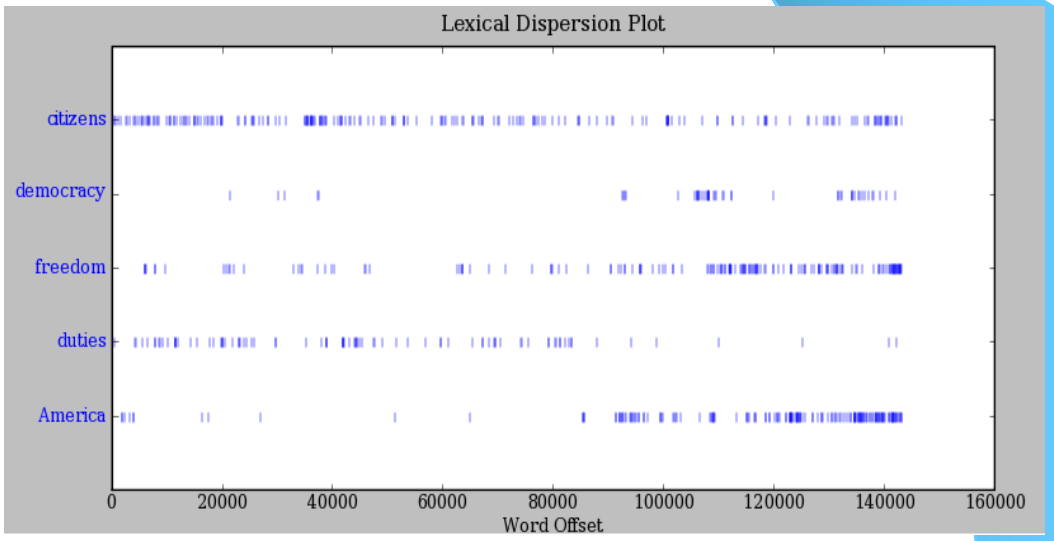


Web Mining



After 9-11: “We”> “I”

3. Text mining:



Multilingual Parallel Corpus mining

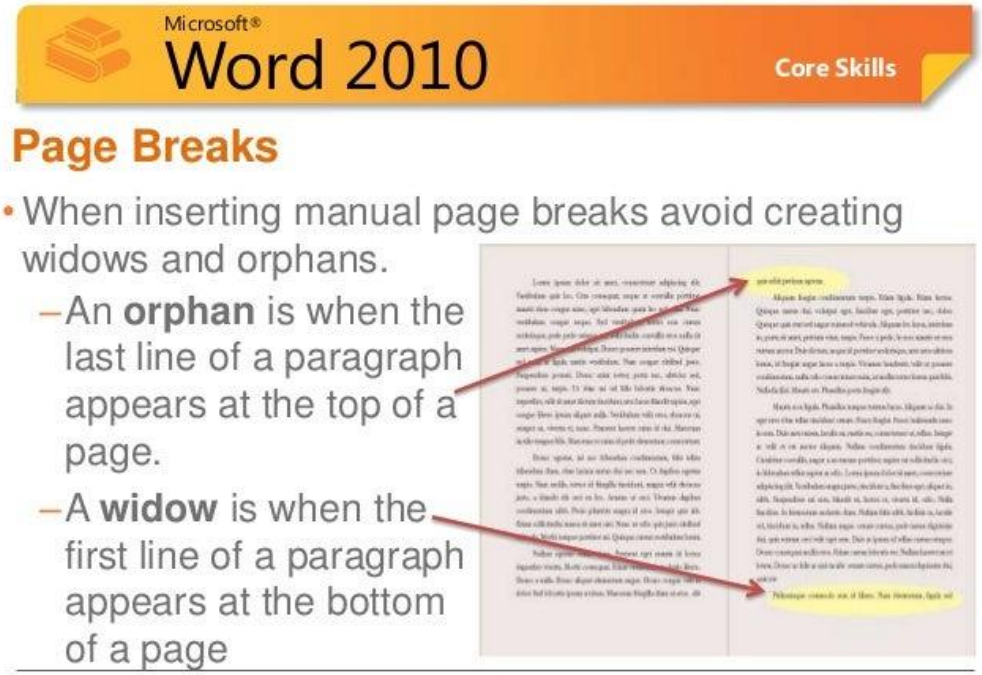
- English: Mother, mom, mum (Brit), mama,
- French: mère, maman; Chinese: 媽媽/mama/; Korean: 엄마
- Bengali: মা/ma/; Thai: แม่/ Mæ/; Khmer: ម៉ាក់/meak/; Hindi: माँ /maan/
- German: Mama, Mutter ;Italian: mamman; Portuguese: mamãe; Spanish: mamá; Swedish, Latvian, Norwegian,…: mamma;
- Danish: mor; Malay: mak; Myanmar:အမေ /a may/;
- Greek: μαμά ; Hebrew: אִמָּא ; Latin: mater
- Russian, Ukrainian, Belarusian, Albanian, Bulgarian, Hungarian, Polish, Dutch, Indonesian, Romanian, Serbian,…: mama;
- Vietnamese: mẹ (Nor) + má (Sou) => mạ (Mid)
- Why?
- */m/ is the easiest phoneme to pronounce for a newborn to pronounce the first word in life.*

Dictionary Mining

- Discover: “The shorter the word, the more common”:
- Ex: “talk”>”communicate”; “nói” > “phát ngôn”
- In Vietnamese dictionary: a hypothesis:
- /ch-v/ => “unstable status”: *chênh vênh*, *chạng vạng*, *chóng vánh*, *chật vật*, *chót vót*, *chới với*, ...
- /e/ => “narrow”: *hẹp*, *dẹp*, *bẹp*, *xẹp*, *lép*, *tép*, *dép*, *hẻm*, *kẹp*, ...
- Is it right ? => mining, calculating statistics.

Mining: English computer terms

- software, hardware, firmware; mother board, parent/child directory, widow, orphan, friend function, sibling disk
- mouse, bug, web, virus; bullet, cylinder, folder, spaghetti-code, wall-paper, stack
- => philosophy: “popular”: informal, avoid formalities, avoid academics (~~dur/dure~~->matériel, ~~doux/douce~~ -> logiciel); close to family, life.



Mining: MultiLingual Parallel Corpus (50L)

<en>We are learning a language.</en>

<fr>Nous apprenons une langue.</fr>

<cn>我们 学习 一门 语言。 </cn>

<ja>言語を 習います。</ja>

<ko>우리는 언어를 배우고 있어요. </ko>

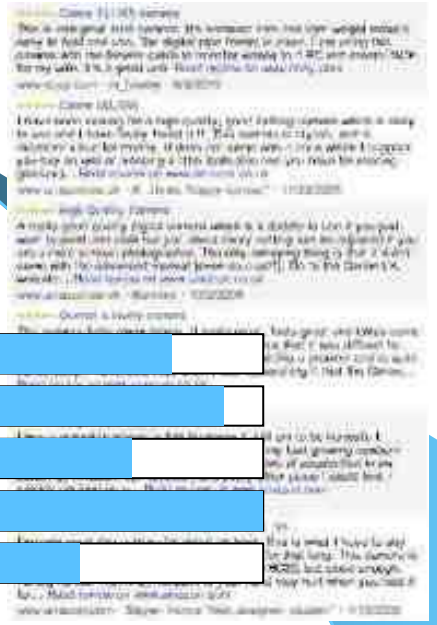
<de>Wir lernen eine Sprache. </de>

<ru>Мы учим язык.</ru>

<eo>Ni lernas lingvon.</eo>

<vi>Chúng ta học một ngôn ngữ.</vi>

Information Extraction & Sentiment Analysis



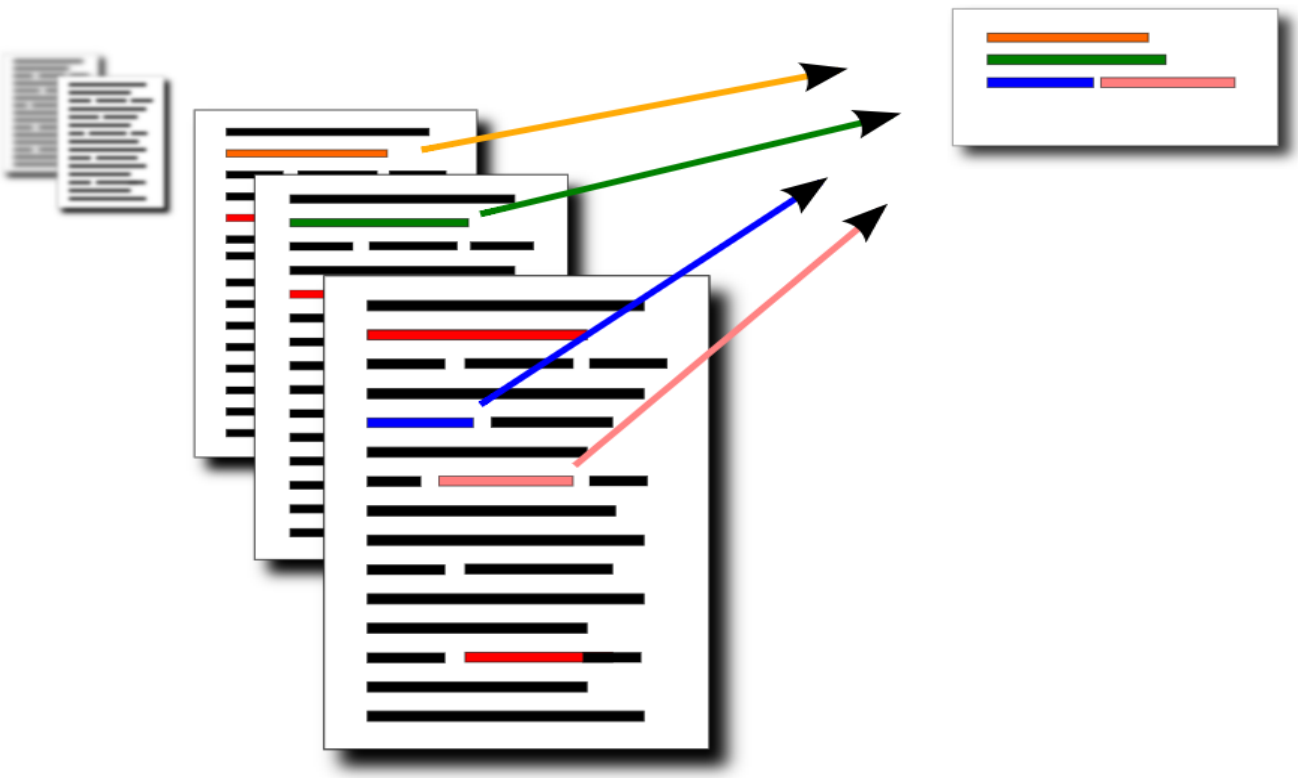
Attributes:

- 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

4. Text summarization:



17 tuổi kiếm triệu USD nhờ bán công ty cho Yahoo

Ứng dụng đọc tin tức Summly của Nick D'Aloisio vừa được Yahoo mua lại với giá gần 30 triệu USD. Cậu cũng trở thành nhân viên chính thức mảng di động của đại gia công nghệ này.

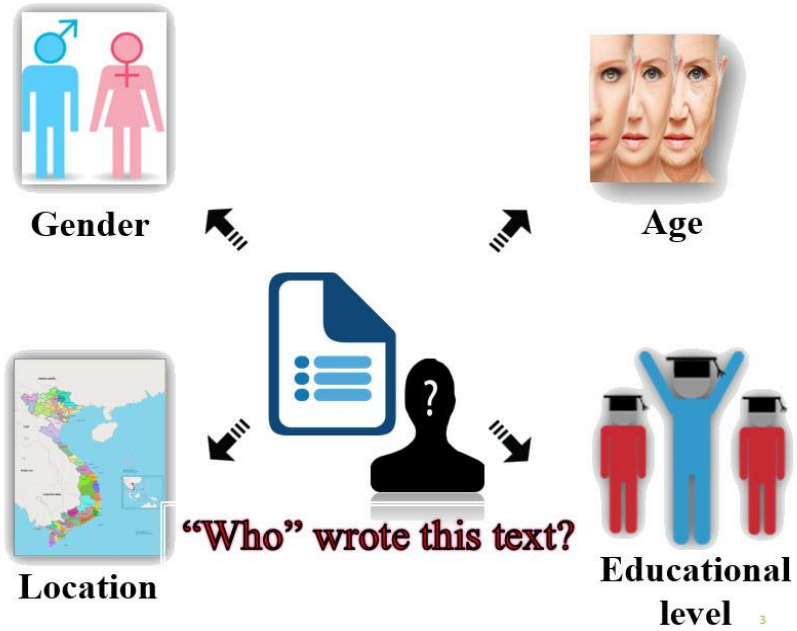
> Con trai Abramovich đầu tư 46 triệu USD vào dầu mỏ

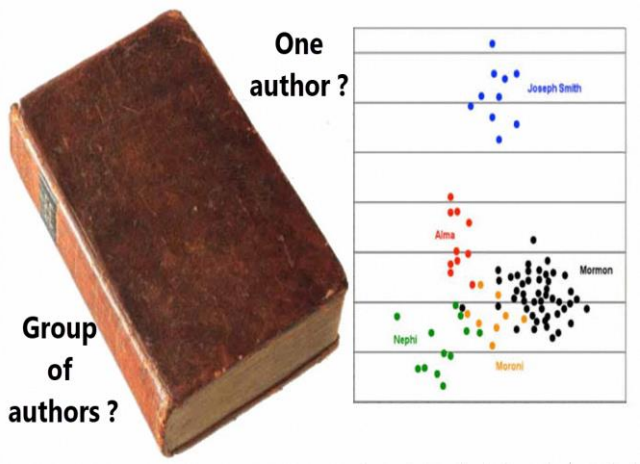
> Nữ giám đốc Việt trong doanh nghiệp của cựu CEO Apple

Nick D'Aloisio, một học sinh 17 tuổi tại Anh, vừa bán công ty của mình - Summly cho Yahoo! với giá gần 30 triệu USD, phần lớn bằng tiền mặt. Hai năm trước, D'Aloisio tạo ra một ứng dụng đọc tin tức cho iPhone có tên "Trimit". Ứng dụng này sẽ lướt qua hàng loạt bài báo online, sau đó tóm tắt chúng thành một đoạn vừa với kích cỡ màn hình điện thoại.

Ban đầu, Trimit không được chú ý nhiều. Nhưng về sau, ứng dụng lọt vào tầm ngắm của tỷ phú Hong Kong - Li Ka Shing. Nhóm đầu tư của ông tại Horizons Ventures đã tài trợ cho D'Aloisio, giúp cậu nâng cấp ứng dụng và đổi tên thành Summly. Sản phẩm còn được quảng cáo bởi rất nhiều người nổi tiếng như Ashton Kucher, Stephen Fry và Yoko Ono.

5. Text stylometry:





Thứ năm, 28/6/2018, 09:28 (GMT+7)

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.

LAW

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

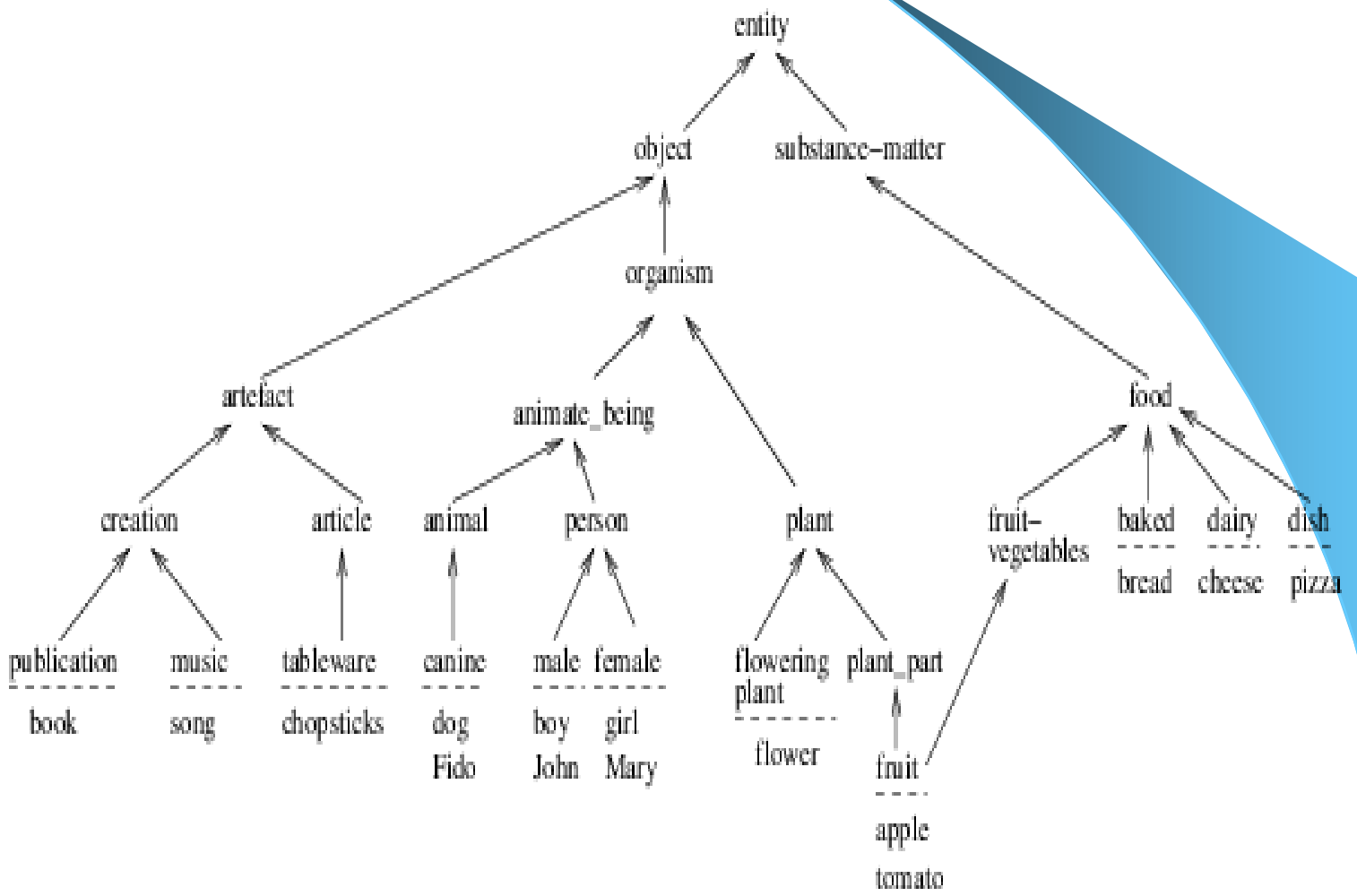
Ted Kaczynski is flanked by federal agents as
Kaczynski is now serving a life sentence in pr

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.

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,

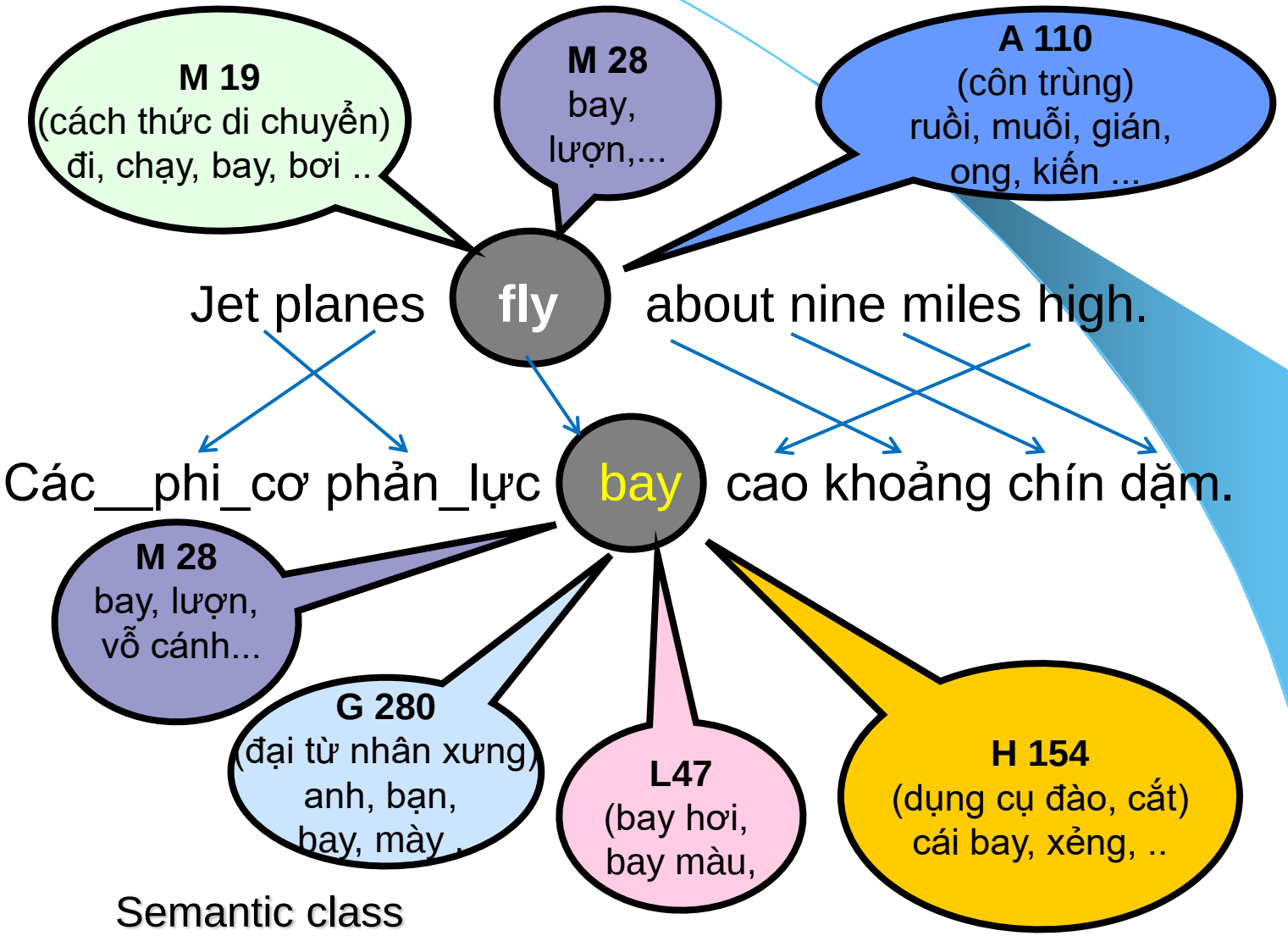
Ex-Math-Prof.
UC Berkeley

6. Text similarity:

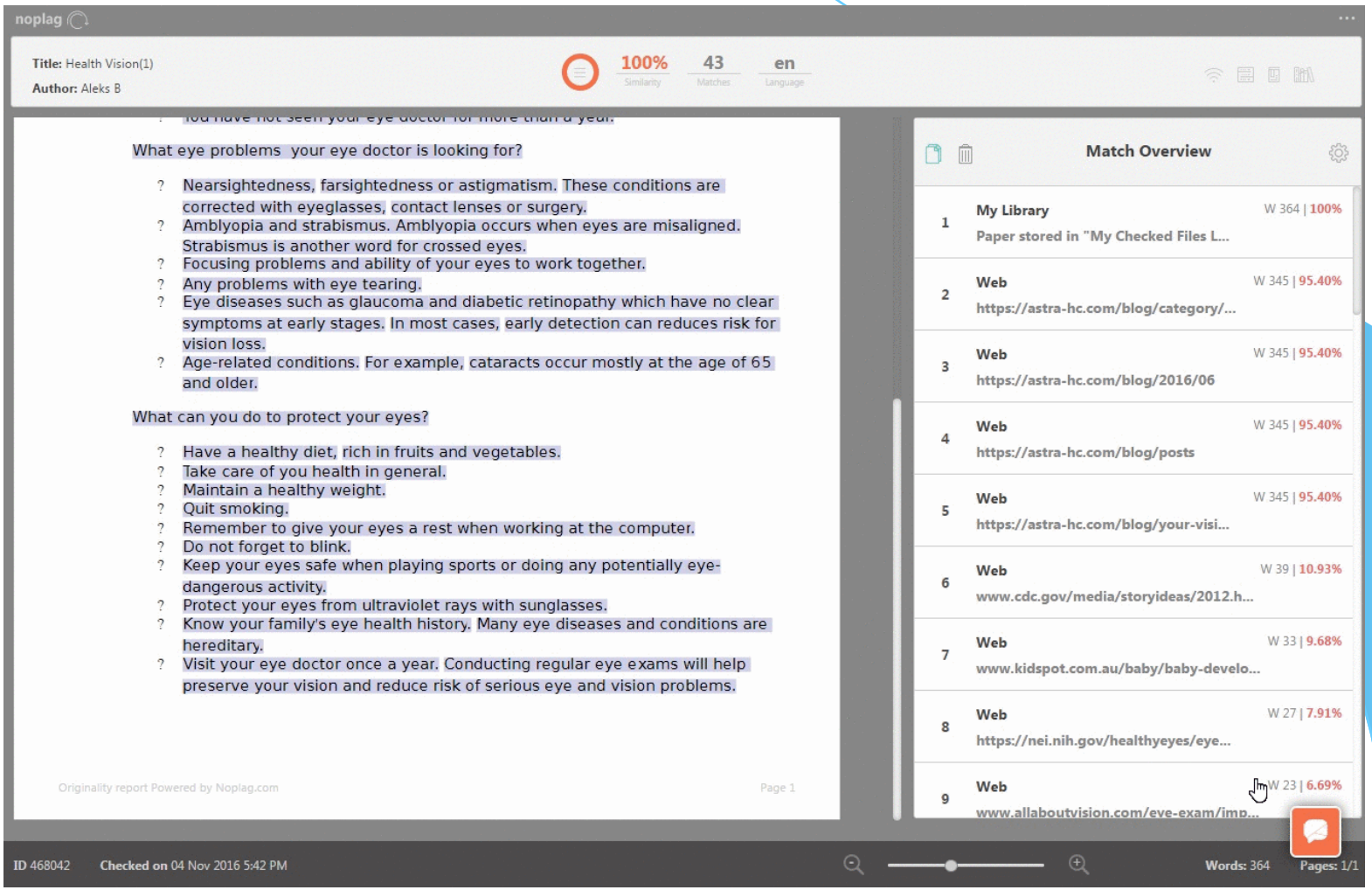


Hyponym/Hypernym of “car” in WordNet

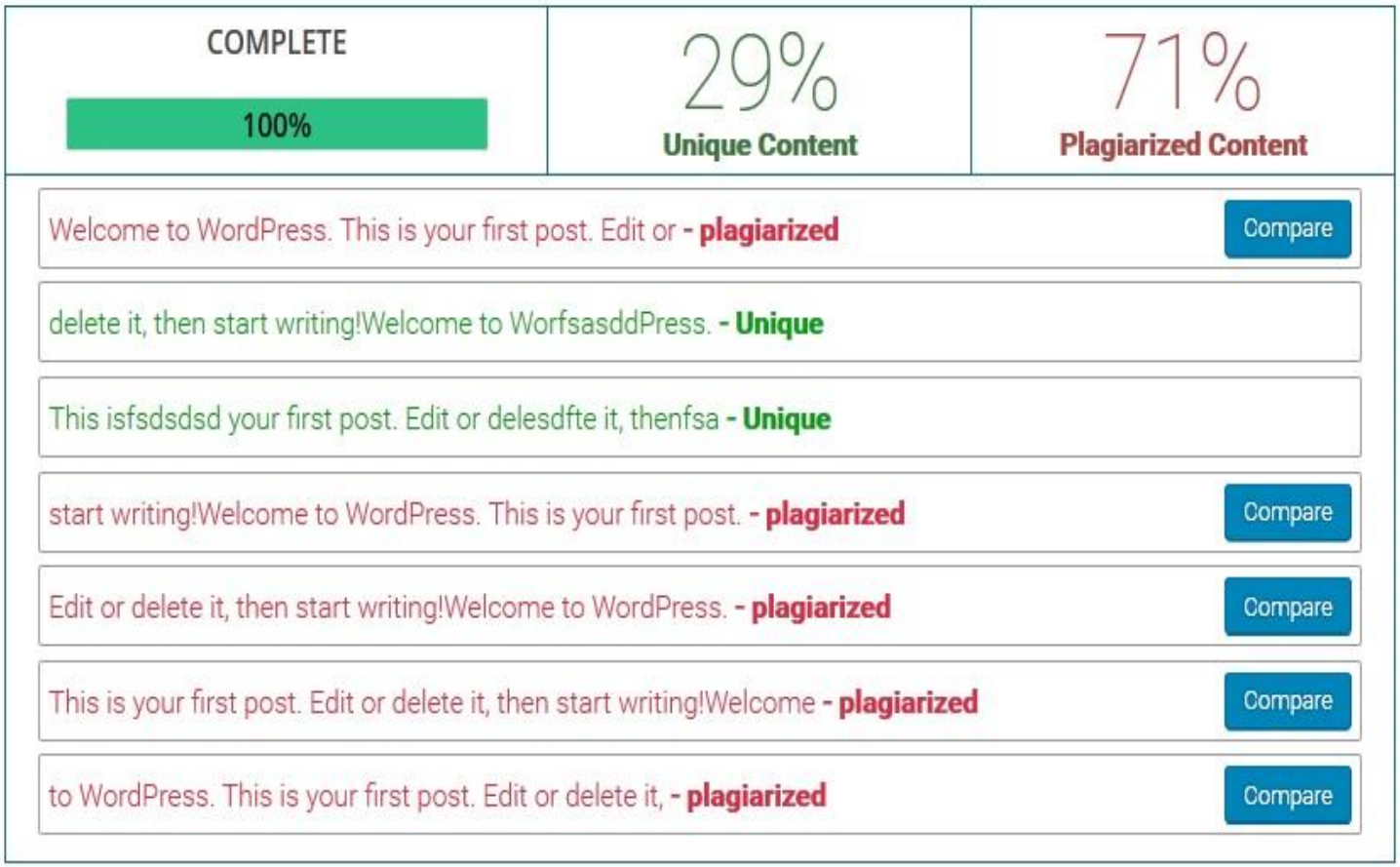
- S: (n) **car**, auto, automobile, machine, motorcar
  - *direct hyponym* / *full hyponym*
  - *part meronym*
  - *domain term category*
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) motor vehicle, automotive vehicle
      - S: (n) self-propelled vehicle
        - S: (n) wheeled vehicle
          - S: (n) vehicle
            - S: (n) conveyance, transport
              - S: (n) instrumentality, instrumentation
                - S: (n) artifact, artefact
                - S: (n) whole, unit
                - S: (n) object, physical object
                - S: (n) physical entity
                - S: (n) entity


M 19
(cách thức di chuyển)
đi, chạy, bay, bơi ..
M 28
bay,
lượn,...
A 110
(côn trùng)
ruồi, muỗi, gián,
ong, kiến ...
Jet planes
fly
about nine miles high.
Các__phi_cơ phản_lực
bay
cao khoảng chín dặm.
M 28
bay, lượn,
vỗ cánh...
G 280
(đại từ nhân xưng)
anh, bạn,
bay, mày ,
L47
(bay hơi,
bay màu,
H 154
(dụng cụ đào, cắt)
cái bay, xẻng, ..
Semantic class

Plagiarism detection



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)

Word frequencies

Rank	Word	f
1	the	1.3712
2	of	1.7254
3	be	1.7322
4	and	1.8024
5	a	1.8101
6	to	1.8224
7	in	1.9179
8	have	2.1294
9	it	2.1578
10	that	2.1992
11	he	2.2107
12	you	2.2921
13	on	2.3084
14	with	2.3312
15	for	2.3710
16	his	2.4079
17	as	2.4162
18	at	2.4193
19	do	2.4428
20	not	2.5276

Ord.	Mot	f
1	de	1.2661
2	la	1.4824
3	l'	1.5739
4	les	1.6347
5	et	1.6486
6	le	1.6668
7	des	1.6971
8	à	1.7145
9	d'	1.7698
10	en	1.8199
11	du	1.9092
12	un	1.9670
13	une	1.9806
14	est	2.0354
15	dans	2.0644
16	que	2.0777
17	qui	2.1059
18	par	2.1656
19	il	2.1686
20	pour	2.1767

号	字	f
1	的	1.4198
2	是	1.7300
3	不	1.7832
4	我	1.8229
5	一	1.8300
6	有	1.8756
7	大	1.9585
8	在	2.0009
9	人	2.0315
10	了	2.0571
11	中	2.1139
12	到	2.1177
13	资	2.1878
14	要	2.2207
15	以	2.2375
16	可	2.2384
17	这	2.2413
18	个	2.2650
19	你	2.2737
20	会	2.2836

Word frequencies

順位	語	f
1	の	1.1763
2	を	1.4490
3	は	1.4556
4	に	1.5124
5	が	1.5209
6	た	1.5215
7	と	1.6344
8	て	1.6484
9	で	1.7151
10	も	2.0045
11	いる	2.1238
12	日	2.1534
13	する	2.1873
14	から	2.2095
15	ない	2.2205
16	こと	2.3060
17	だ	2.3399
18	など	2.3865
19	人	2.4126
20	年	2.4189

ранг	слово	f
1	и	1.4412
2	не	1.6902
3	он	1.7219
4	на	1.7822
5	я	1.8174
6	что	1.8697
7	тот	1.9466
8	быть	1.9488
9	с	1.9535
10	а	2.0177
11	весь	2.0700
12	это	2.0965
13	как	2.1447
14	она	2.1575
15	по	2.2354
16	но	2.2458
17	они	2.2476
18	к	2.2622
19	у	2.3049
20	из	2.3172

Stt	Từ	f
1	và	1.8199
2	của	1.8223
3	có	1.9560
4	các	1.9591
5	là	1.9682
6	một	1.9861
7	được	2.0125
8	không	2.0317
9	trong	2.0433
10	cho	2.0504
11	đã	2.0826
12	những	2.1467
13	với	2.1483
14	người	2.1599
15	ở	2.2107
16	để	2.2750
17	này	2.2809
18	đến	2.3032
19	vào	2.3090
20	tôi	2.3140

Vietnamese Word-usage Frequency

No	Word	POS (en)	f
1	của	Cm (of)	1.820
2	và	Cp (and)	1.822
3	các	Nq (+PLR)	1.956
4	có	Ve (have)	1.959
5	là	Vc (tobe)	1.968
6	trong	Cm (in)	1.986
7	một	Nq (one)	2.012
8	đã	R (+PST)	2.031
9	những	Nq (+PLR)	2.043
10	không	R (no,not)	2.050

Table 6. VN word frequency

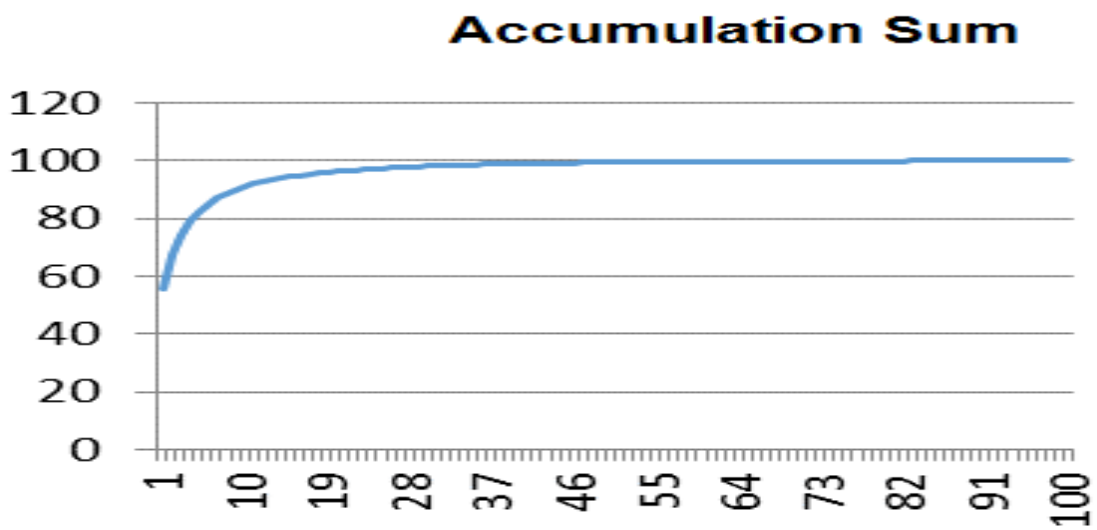
Rank	Word	Eng	POS	f
..	…			..
14	người	man	Nn	2.160
25	nhiều	many	Aa	2.210
27	năm	year	Nt	2.314
30	ngày	day	Nt	2.401
31	làm	do	Vv	2.423
32	phải	must	Vm	2.436
34	ông	you	Pro	2.464
36	theo	follow	Vv	2.530
43	việc	thing	Nn	2.611
53	có thể	able	Vv	2.660

Legend: Cm: prep; Cp: conj; Nq: quatifier, Ve: Verb-exist; Vc: copula; R: adverb; Nn: Noun, Vv: Verb, Aa: Adj, M: Modifier.

Vietnamese Word Frequency (homonym)

Rank	Word	Eng	POS	f
3,775	của	wealth	Nn	4.6789
368	và	then	M	3.4268
20,793	và	shovel	Vv	6.1384
39,212	các	pay.extra	Vv	6.7405
3,224	có	(particle)	M	4.5731
103	có	exist	R	2.9803
19,385	là	iron	Vv	6.0415
5,290	là	being	Cs	4.9209
143	là	as	Cp	3.0857
1,749	là	(particle)	M	4.1842
186	tốt	good	Aa	3.1813
25,154	tốt	soldier	Nn	6.4394

Vietnamese Word-usage Frequency



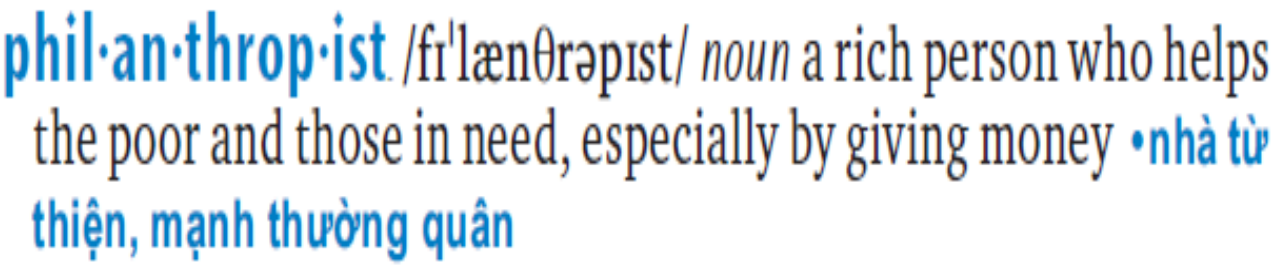
- Top 10% of most popular word types (3,400 words)
  - Cover 90% of word tokens in texts.
- Only teaching top-3,400 word types.
  - Able to read 90% content of Vietnamese texts.

Applications of Text readability

- Compiling texts (on topics) and ranked (easy, normal, difficult).
- Textbook editors: select appropriate texts on topic and learner's level (avoid subjective selections).
- ✓ Questions in examination: CEFR (A1,A2,..C2)
- ✓ Recruiting candidates of reporters.
- ✓ Writing user-manual of pesticide for peasants.
- ✓ Writing user-manual installation for workers
- ✓ To protect customers in healthcare contracts from disputing.
- ✓ Writing definitions in dictionaries.


Education
ĐẠI TRƯỜNG PHÁT
RA MẮT TỪ ĐIỂN OALD ANH VIỆT
& GIẢI PHÁP DẠY VÀ HỌC TIẾNG ANH
Hà Nội, ngày 19 tháng 03 năm 2015
Oxford
Advanced
Learner's
Dictionary
with Vietnamese Translation
Oxford Writing Tutor
Richmond
EnglishCentral

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



- 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

8. Text translation:

8a. Machine Translation

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.

Rmsfield and Cheney insist on keeping the American forces in Iraq

Called American Defense Minister Donald Rmsfield 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.

Voice translation

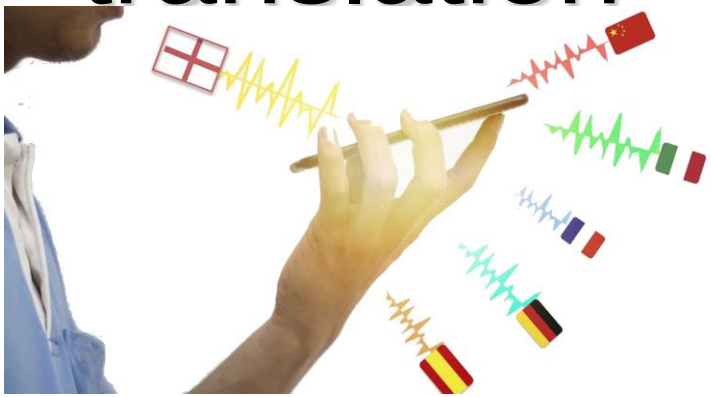
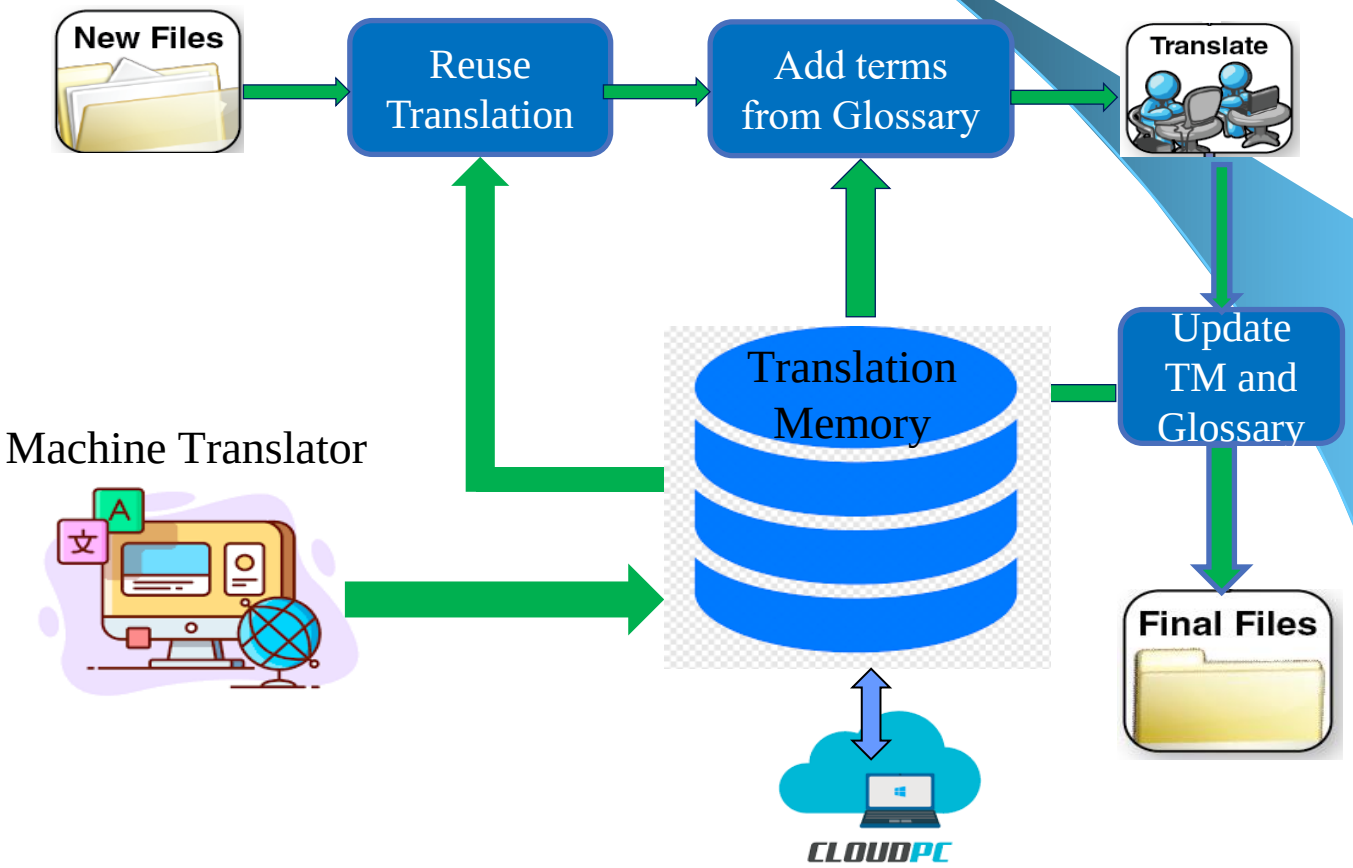


Image translation

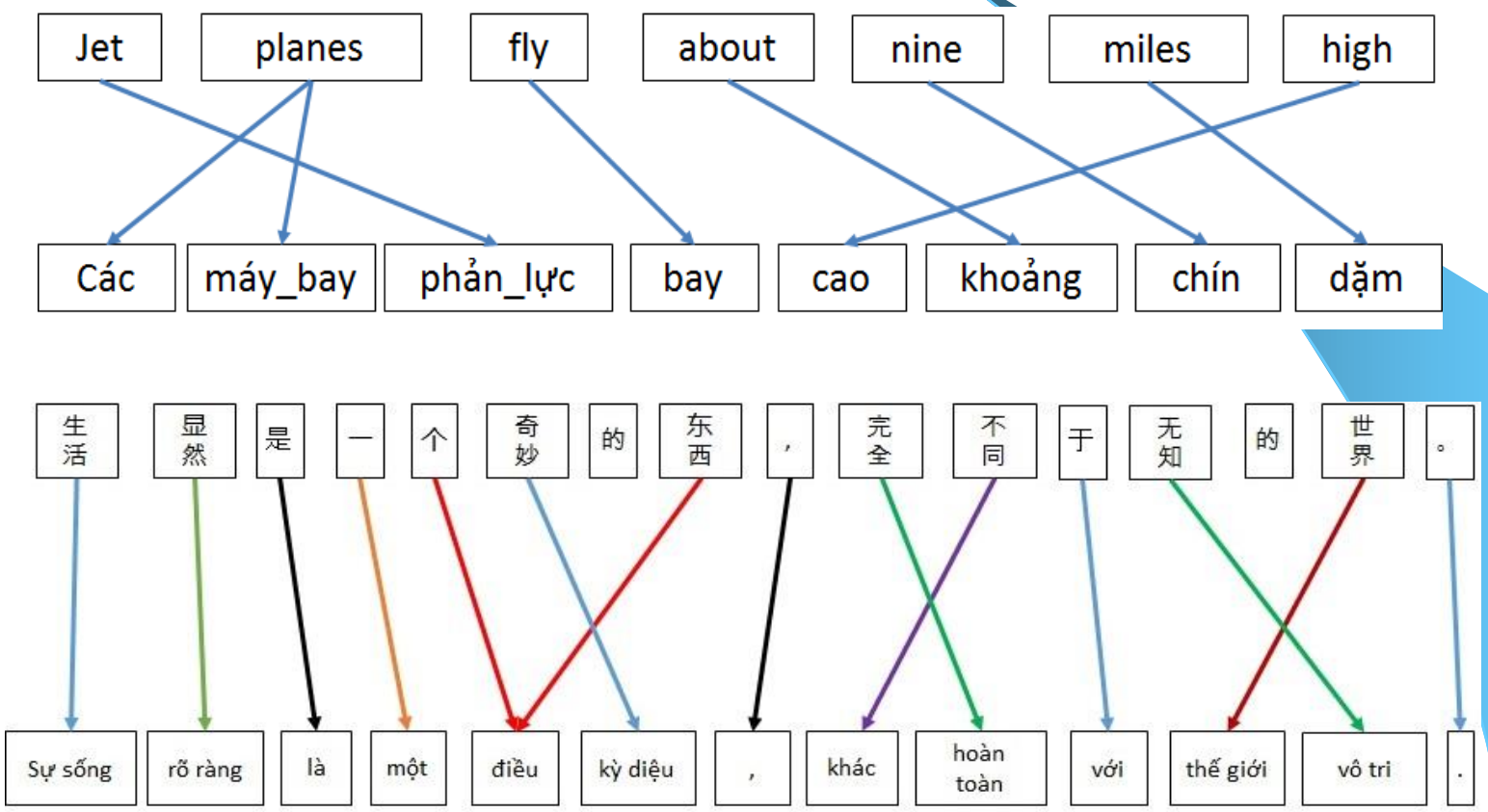
8a. CAT

Computer Assisted Translation

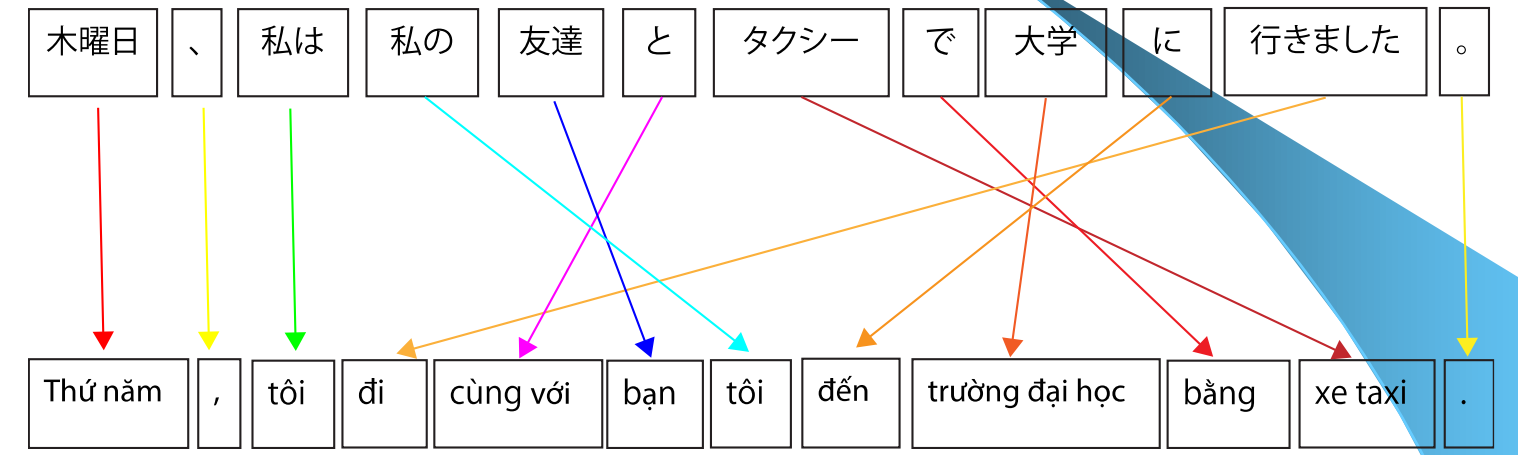
Translation Memory, Glossary

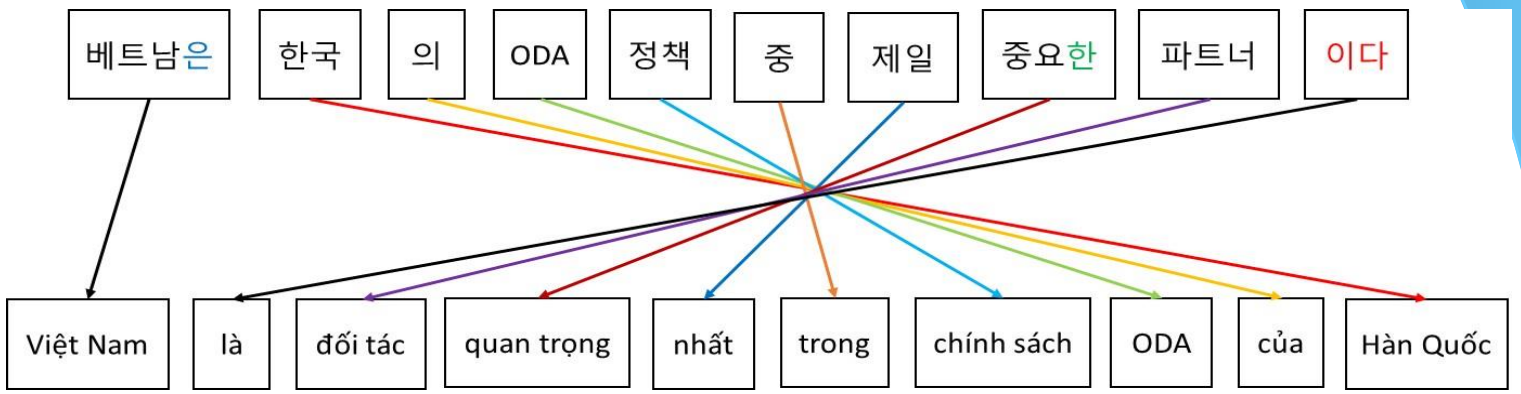


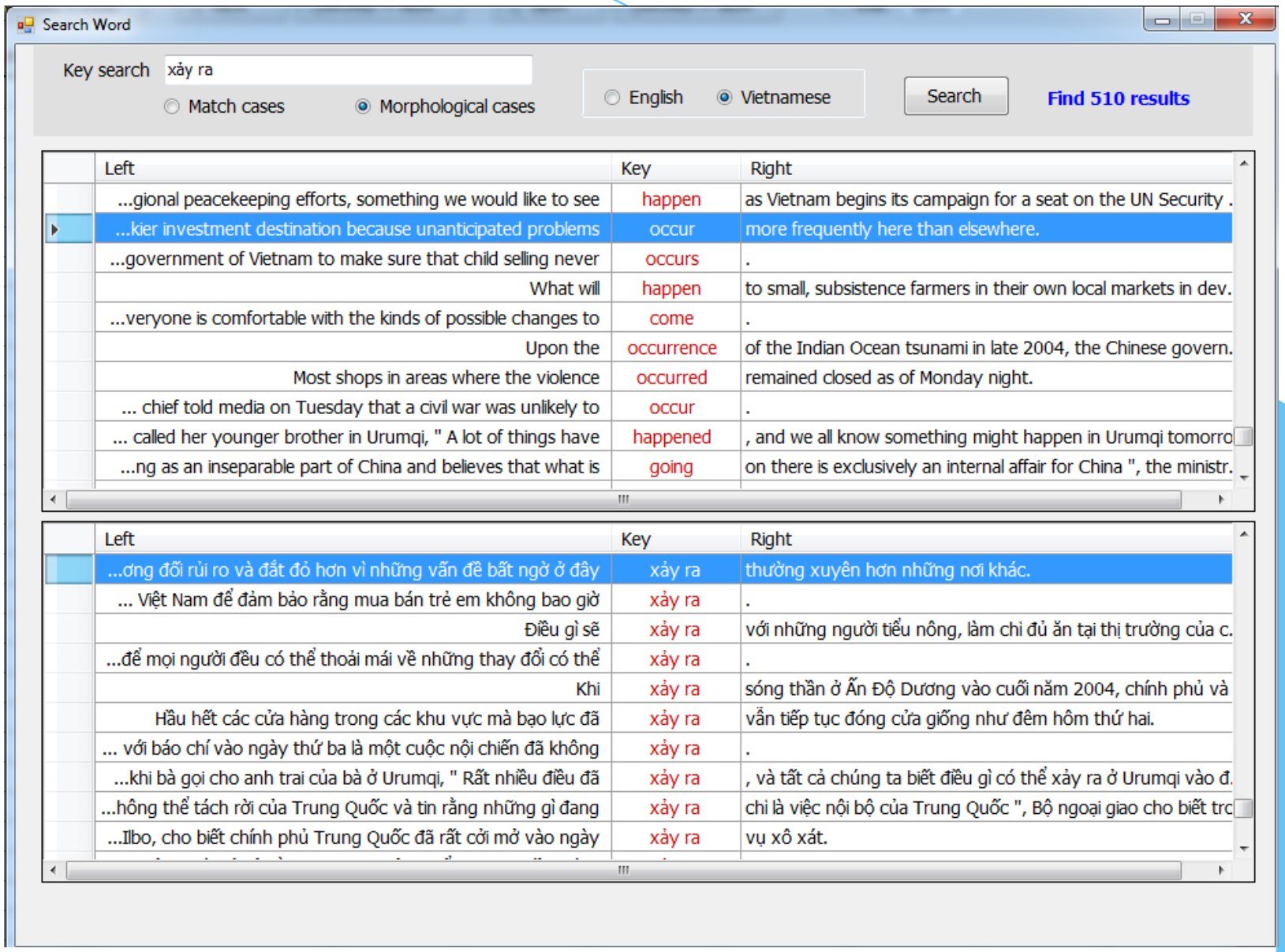
9. Parallel text processing: Contrastive Linguistics

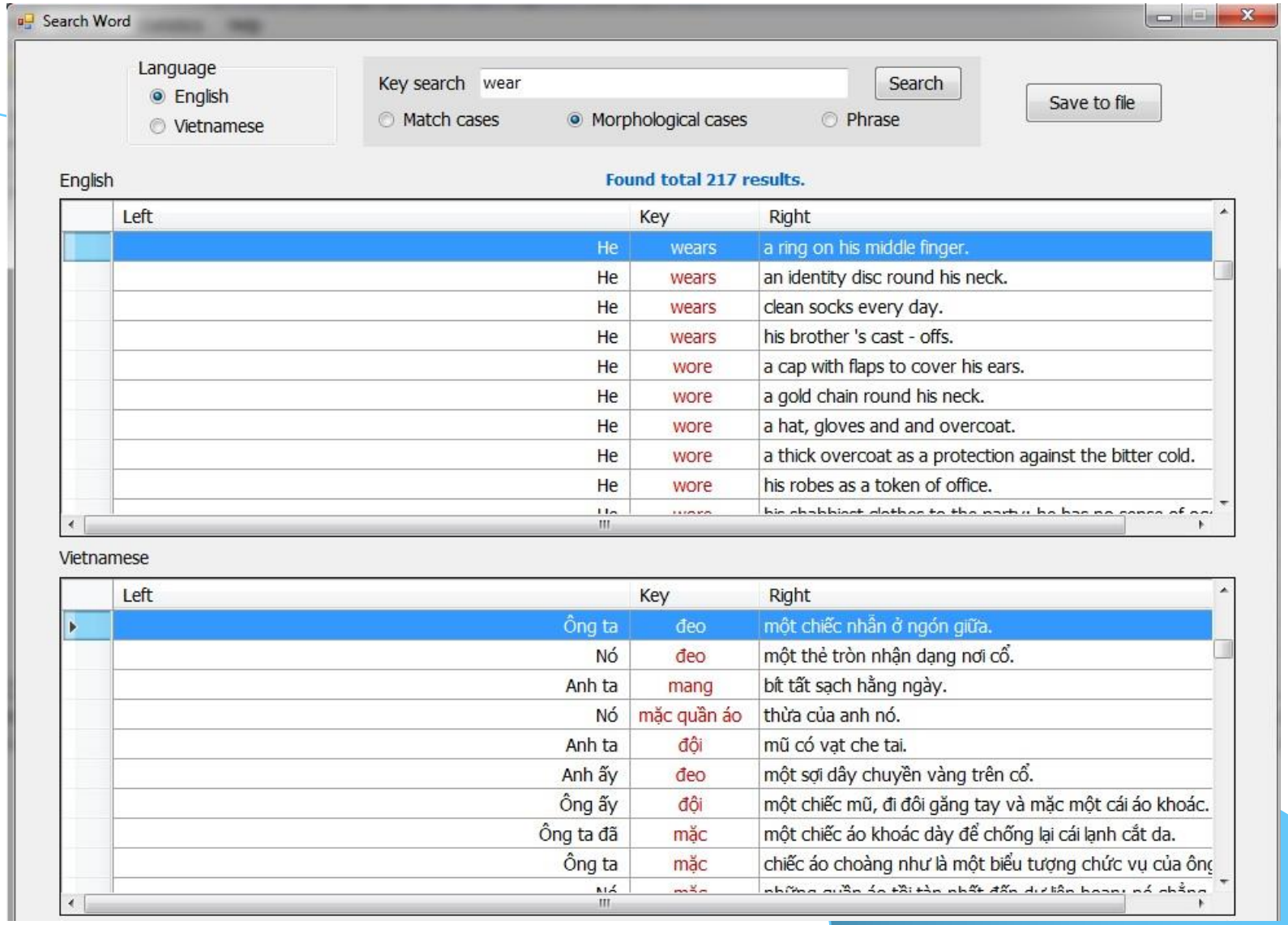


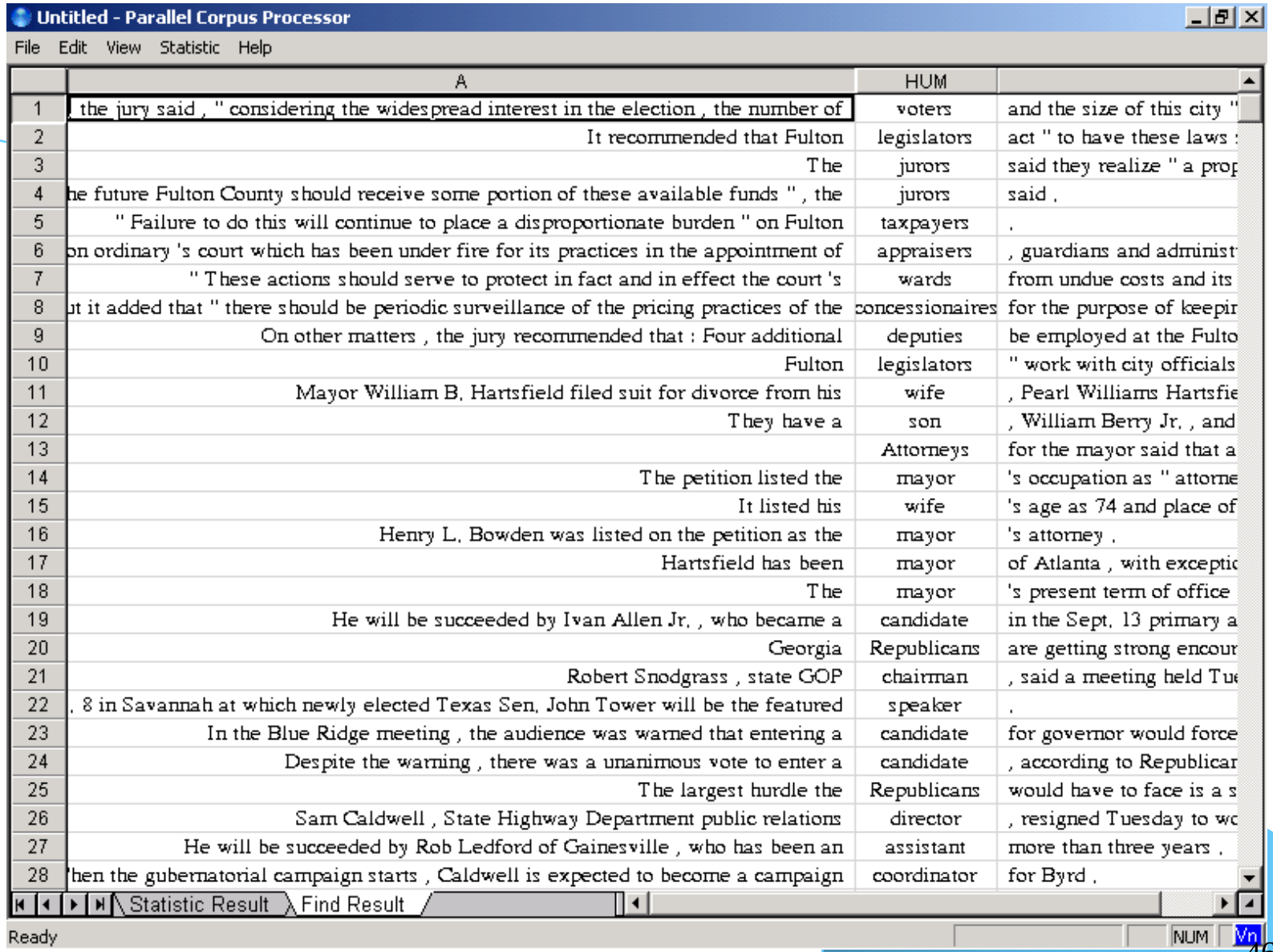
Parallel Corpora

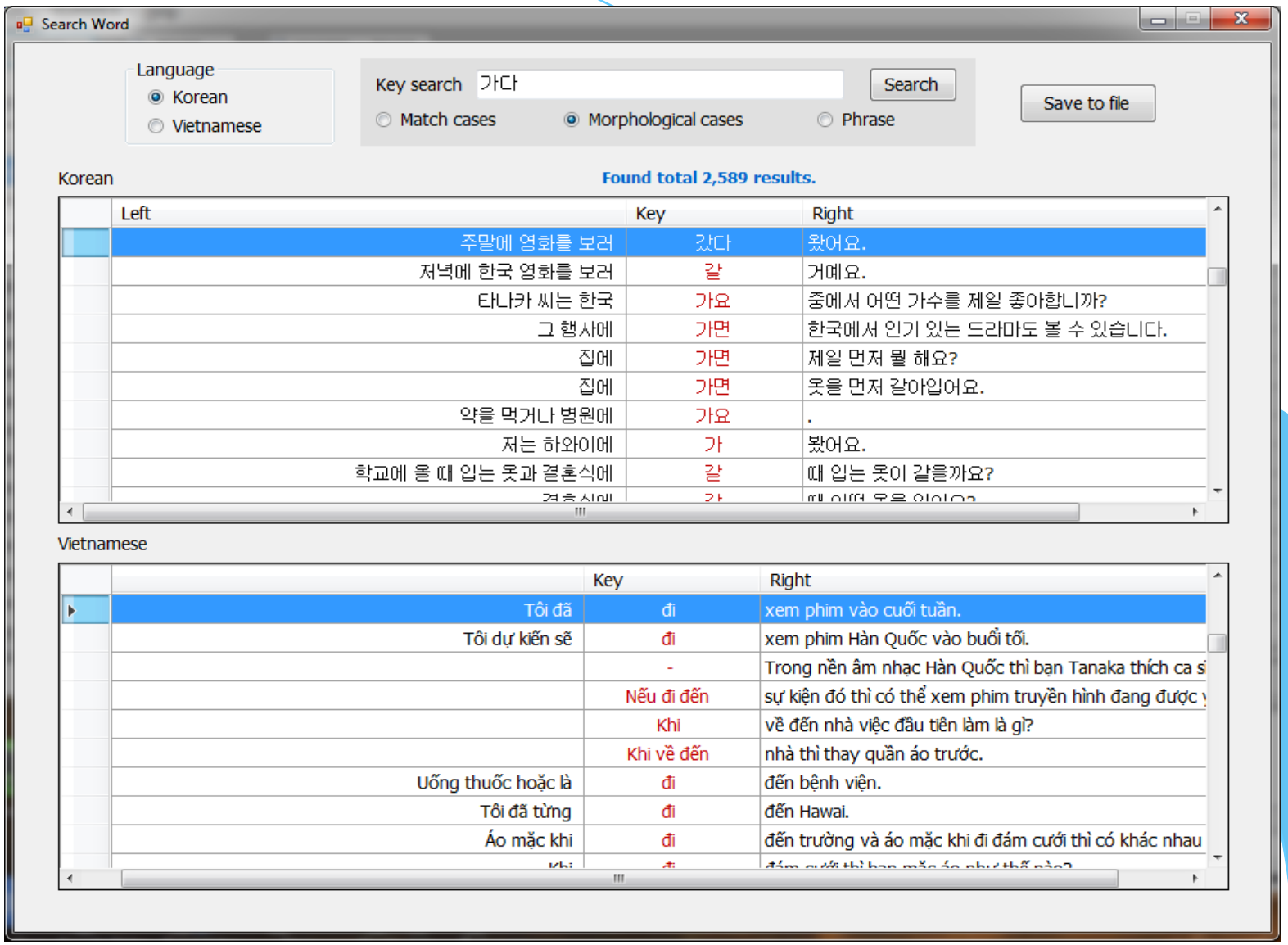




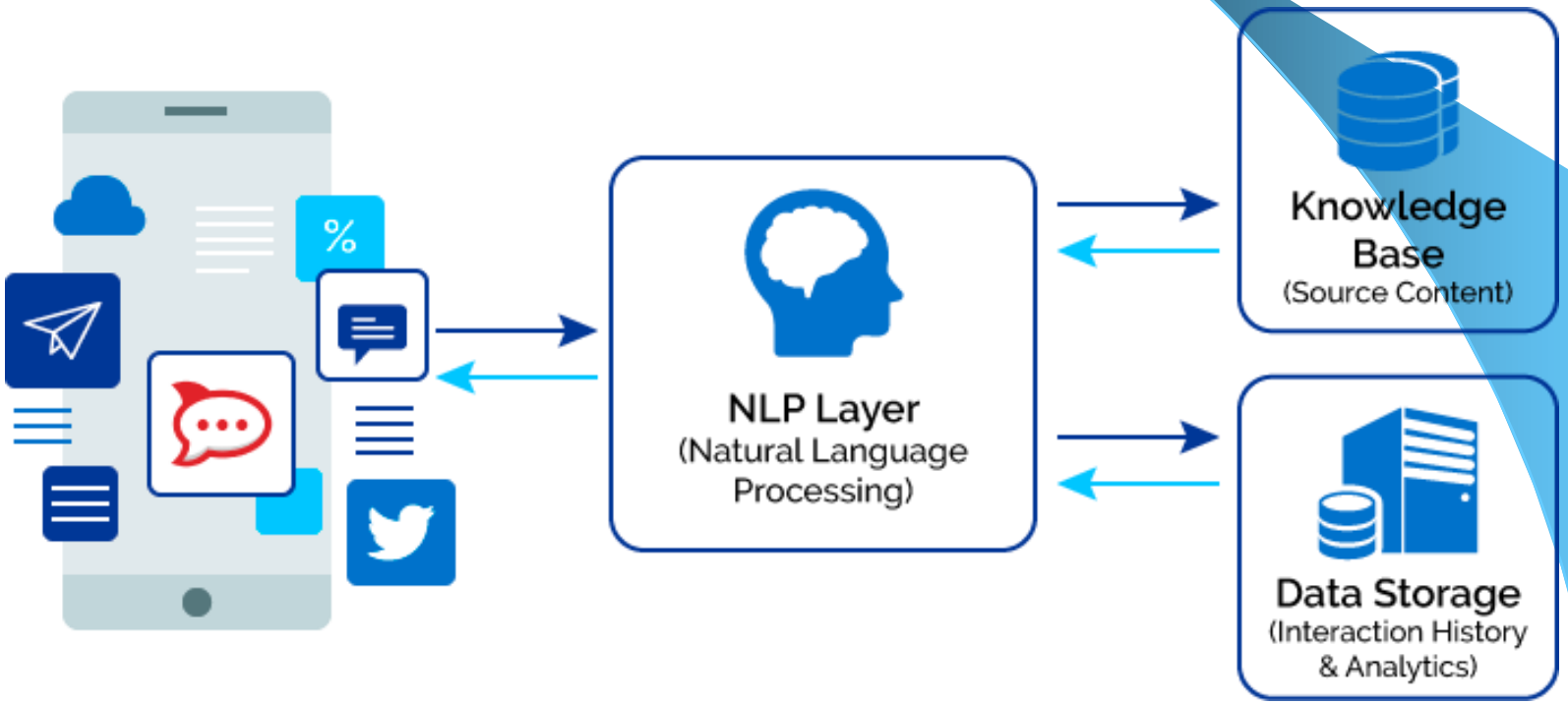

Search Word
Key search
xảy ra
Match cases
Morphological cases
English
Vietnamese
Search
Find 510 results
Left
Key
Right
...gional peacekeeping efforts, something we would like to see
happen
as Vietnam begins its campaign for a seat on the UN Security .
...kier investment destination because unanticipated 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.
...veryone 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 tomorro
...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
...ơng 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 tiểu 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
chỉ là việc nội bộ của Trung Quốc ", Bộ ngoại giao cho biết trc
...Ilbo, cho biết chính phủ Trung Quốc đã rất cởi mở vào ngày
xảy ra
vụ xô xát.


Search Word
Language
English
Vietnamese
Key search
wear
Search
Match cases
Morphological cases
Phrase
Save to file
English
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.
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 thừa 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, đi đôi 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

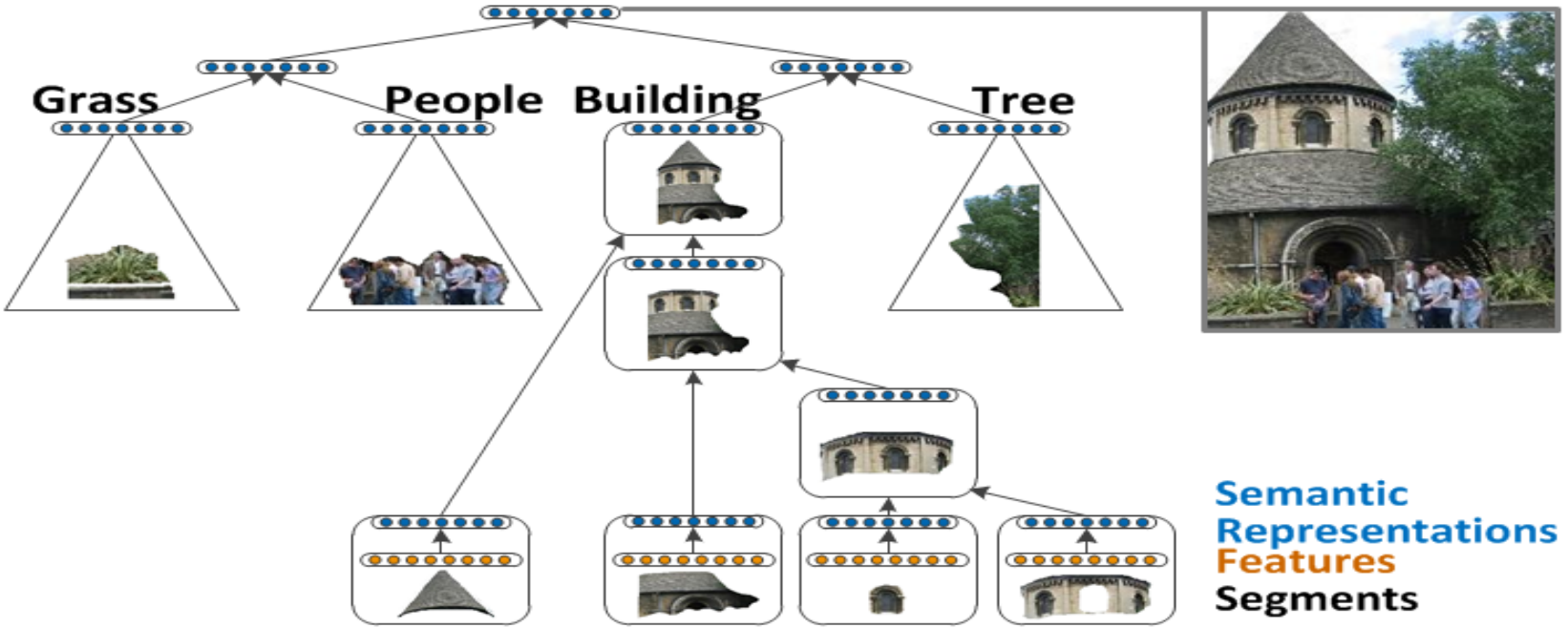

Untitled - Parallel Corpus Processor
File Edit View Statistic Help
A
HUM
1 , the jury said , " considering the widespread interest in the election , the number of voters and the size of this city "
2 It recommended that Fulton legislators act " to have these laws :
3 The jurors said they realize " a 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 ut 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 Attorneys for the mayor said that a
14 The petition listed the mayor 's occupation as " attorne
15 It listed his wife 's age as 74 and place of
16 Henry L. Bowden was listed on the petition as the mayor 's attorney .
17 Hartsfield has been mayor of Atlanta , with exceptic
18 The mayor 's present term of office
19 He will be succeeded by Ivan Allen Jr. , who became a candidate in the Sept. 13 primary a
20 Georgia Republicans are getting strong encour
21 Robert Snodgrass , state GOP chairman , said a meeting held Tu
22 , 8 in Savannah at which newly elected Texas Sen. John Tower will be the featured speaker .
23 In the Blue Ridge meeting , the audience was warned that entering a candidate for governor would force
24 Despite the warning , there was a unanimous vote to enter a candidate , according to Republican
25 The largest hurdle the Republicans would have to face is a s
26 Sam Caldwell , State Highway Department public relations director , resigned Tuesday to wo
27 He will be succeeded by Rob Ledford of Gainesville , who has been an assistant more than three years .
28 hen the gubernatorial campaign starts , Caldwell is expected to become a campaign coordinator for Byrd .
Statistic Result
Find Result
Ready
NUM


Search Word
Language
Korean
Vietnamese
Key search
가다
Search
Match cases
Morphological cases
Phrase
Save to file
Korean
Found total 2,589 results.
Left
Key
Right
주말에 영화를 보러
갔다
왔어요.
저녁에 한국 영화를 보러
갈
거예요.
타나카 씨는 한국
가요
중에서 어떤 가수를 제일 좋아합니까?
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?
Vietnamese
Key
Right
Tôi đã
đi
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 s
Nếu đi đến
sự kiện đó thì có thể xem phim truyền hình đang được
Khi
về đến nhà việc đầu tiên làm là gì?
Khi về đến
nhà thì thay quần áo trước.
Uống thuốc hoặc là
đi
đến bệnh viện.
Tôi đã từng
đi
đến Hawai.
Áo mặc khi
đi
đến trường và áo mặc khi đi đám cưới thì có khác nhau

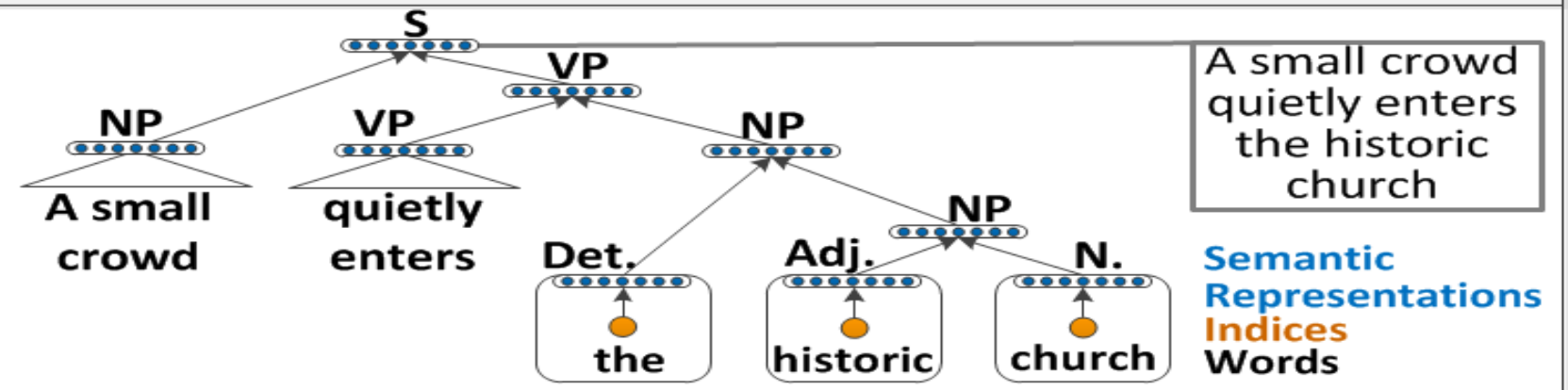
10. Chatbot


%
NLP Layer
(Natural Language Processing)
Knowledge Base
(Source Content)
Data Storage
(Interaction History & Analytics)

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 khố"
- Corpus = Collection of spoken/written text
- Building Corpus: by native-speaker, Master in Applied Linguistics (Computational Linguistics), Data Science.

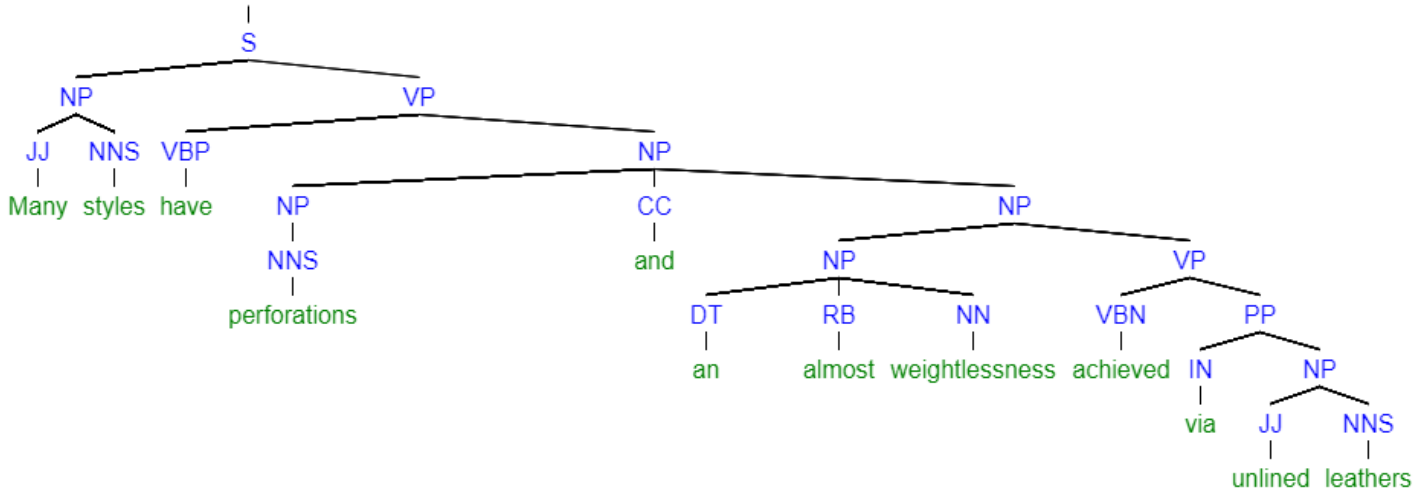
Corpus:

- 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 trục/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 York Times

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?

Á hậu, MC bình thản khai nhận
nhiều lần bán dâm ngàn USD cho
khách 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

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

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> 8/9/13

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

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
Subject: Acquiring Vietnamese treebank

Dear Prof. Dinh Dien,

HyunJeong Choe <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 <AdvancedLinguisticsgmail.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(

to clc

Greetings,

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

Số hóa > Công nghệ Thứ tư, 30/9/2020, 13:00 (GMT+7)

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.

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.

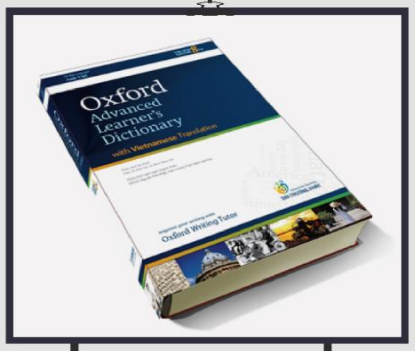
Would it be possible talking with teachers from the Center?

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>
```



Not secure | kimtudien.com.vn

Apps | Tin nhanh VnExpres... | Eximbank | Google | Báo Dân trí – Thôn... | YouTube | tu dien Han Viet | Ngân hàng thương... | Device(Foscam BR

VIETNAMESE

Chuyển tự tự động từ Chữ Nôm sang Chữ Quốc Ngữ

Thể loại của văn bản

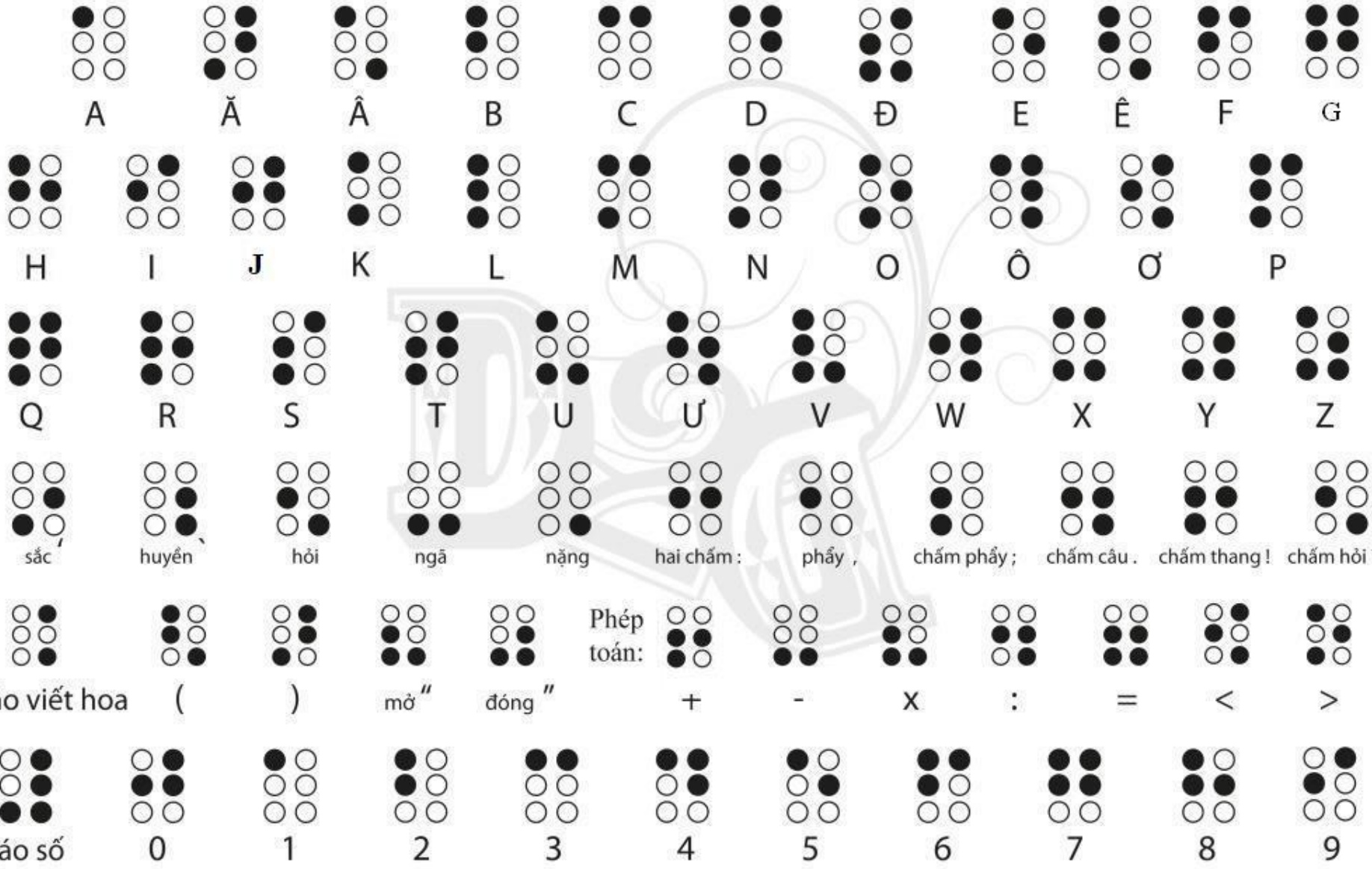
đời sống | văn học | tôn giáo

Văn bản Chữ Nôm

Dịch

ngẫm thay muôn sự bởi trời
trời kia đã bắt làm người có thân
bắt phong trần, phải phong trần
cho thanh cao mới được phần thanh cao
có đâu thiên vị người nào
chữ tài chữ mệnh dồi dào cả hai
có tài mà cậy chi tài
chữ tài liền với chữ tai một vần
đã mang lấy nghiệp vào thân
cũng đừng trách lẫn trời gần trời xa

KÝ HIỆU CHỮ BRAILLE VIỆT NGỮ



Trung tâm dữ liệu đa ngữ

Kim Từ Điển (KMDC) chúng

tôi chuyên sản xuất các

Phần mềm có hỗ trợ tiếng

Việt cho người khiếm thị.

Dan Jurafsky

Language Technology

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

making good progress

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 13th Shanghai International Film Festival...

Information extraction (IE)

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

Party
May 27
add

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

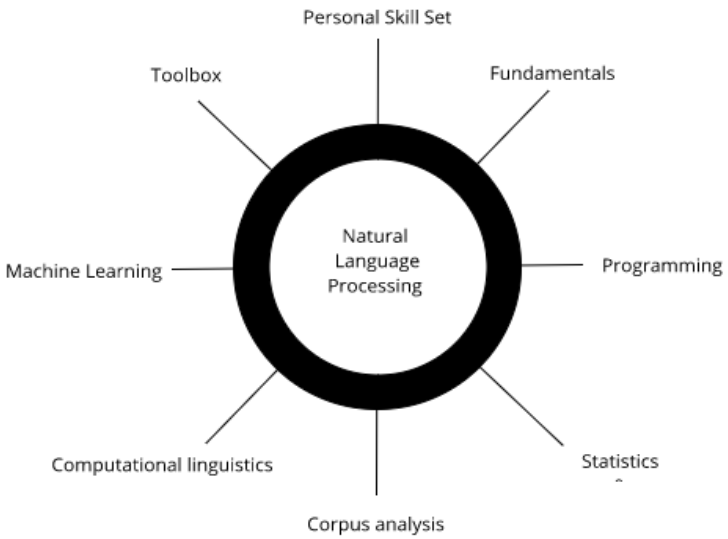
⇨ Economy is good

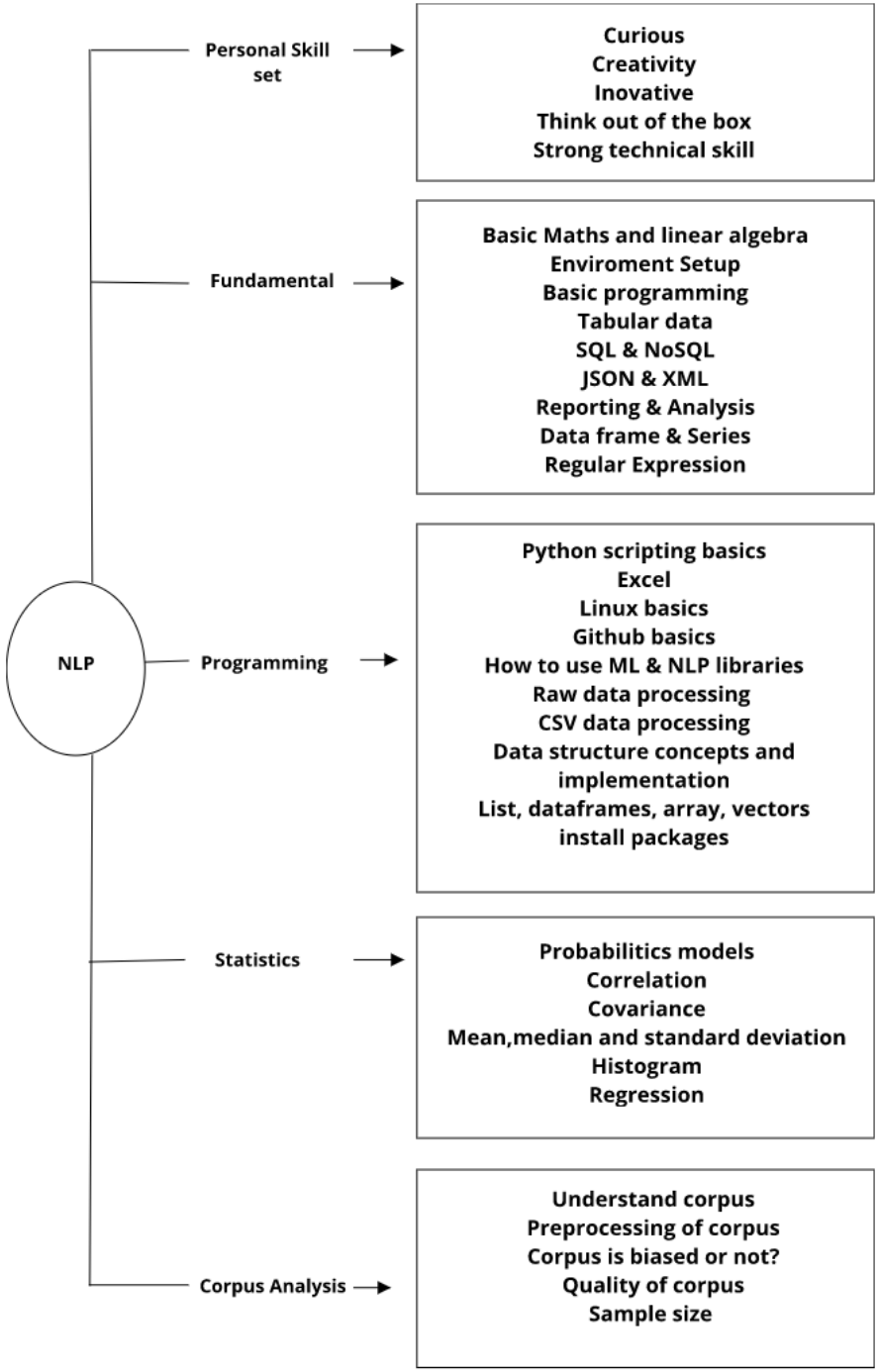
Dialog

Where is Citizen Kane playing in SF?

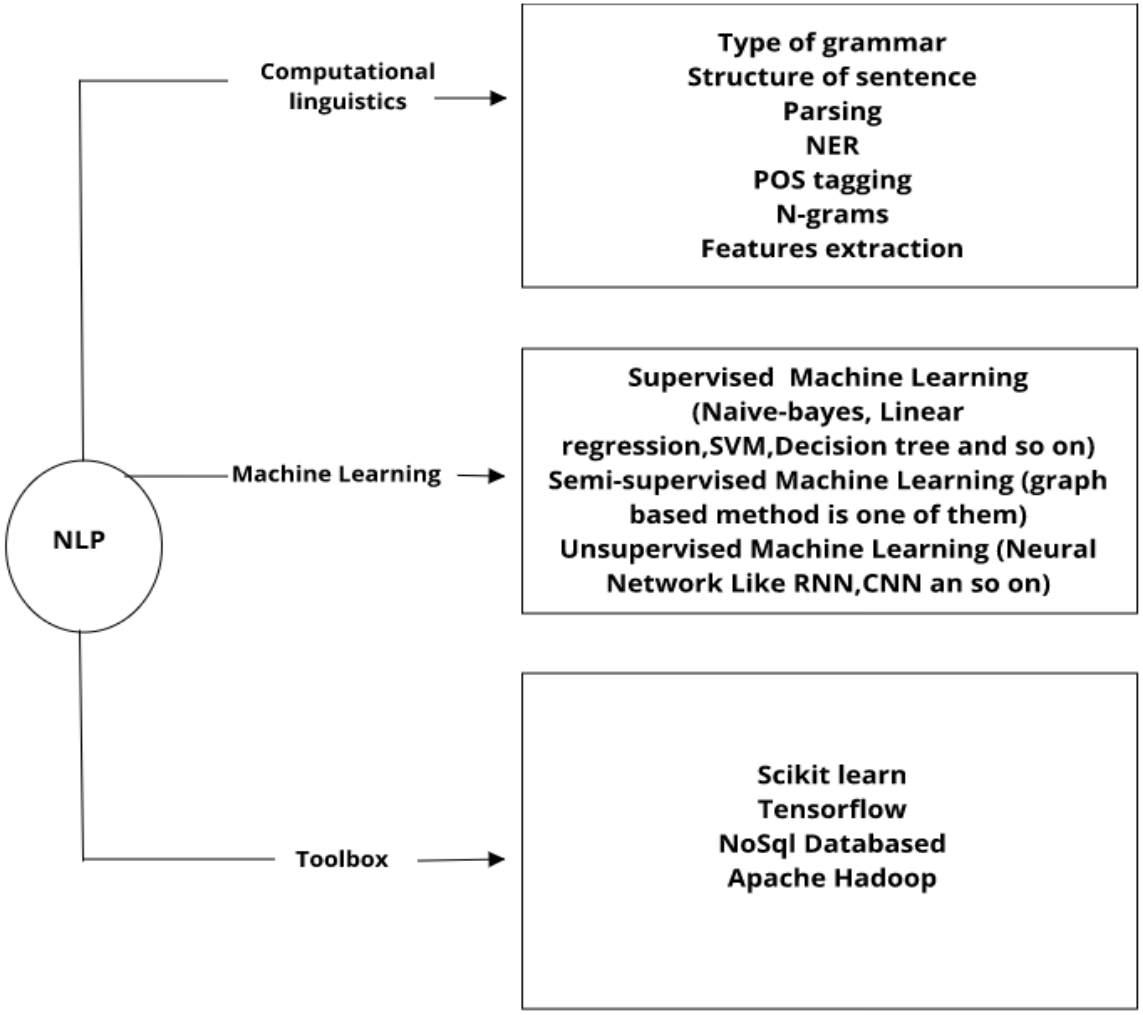
Castro Theatre at 7:30. Do you want a ticket?

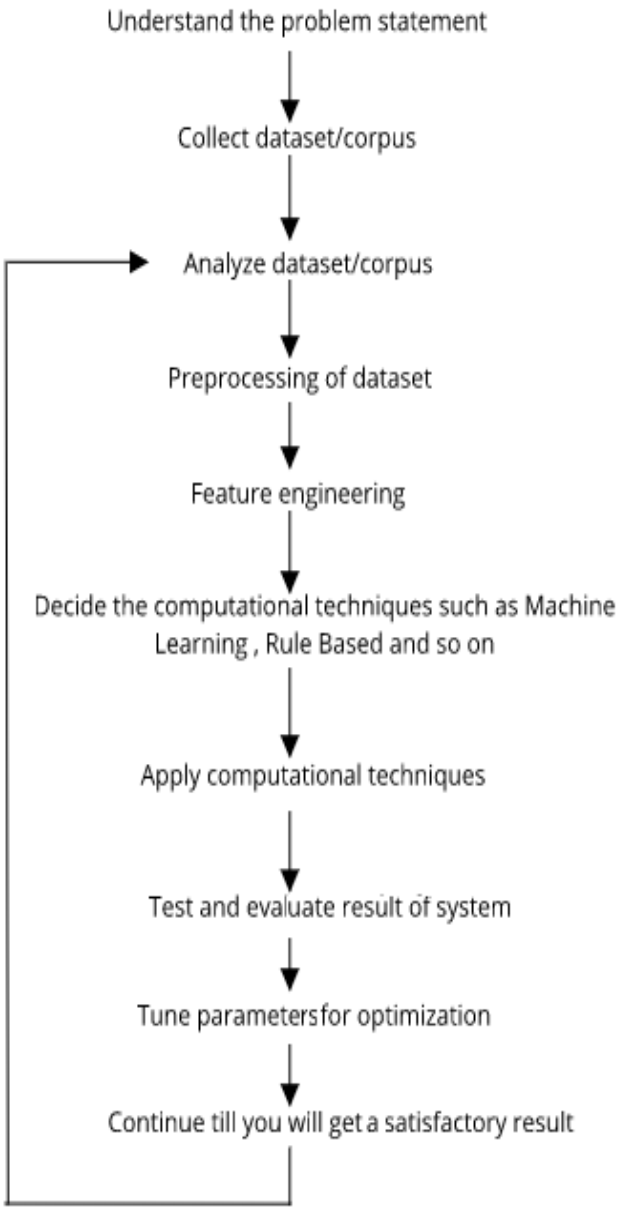
What do you need for NLP?





What will you learn in this NLP course?

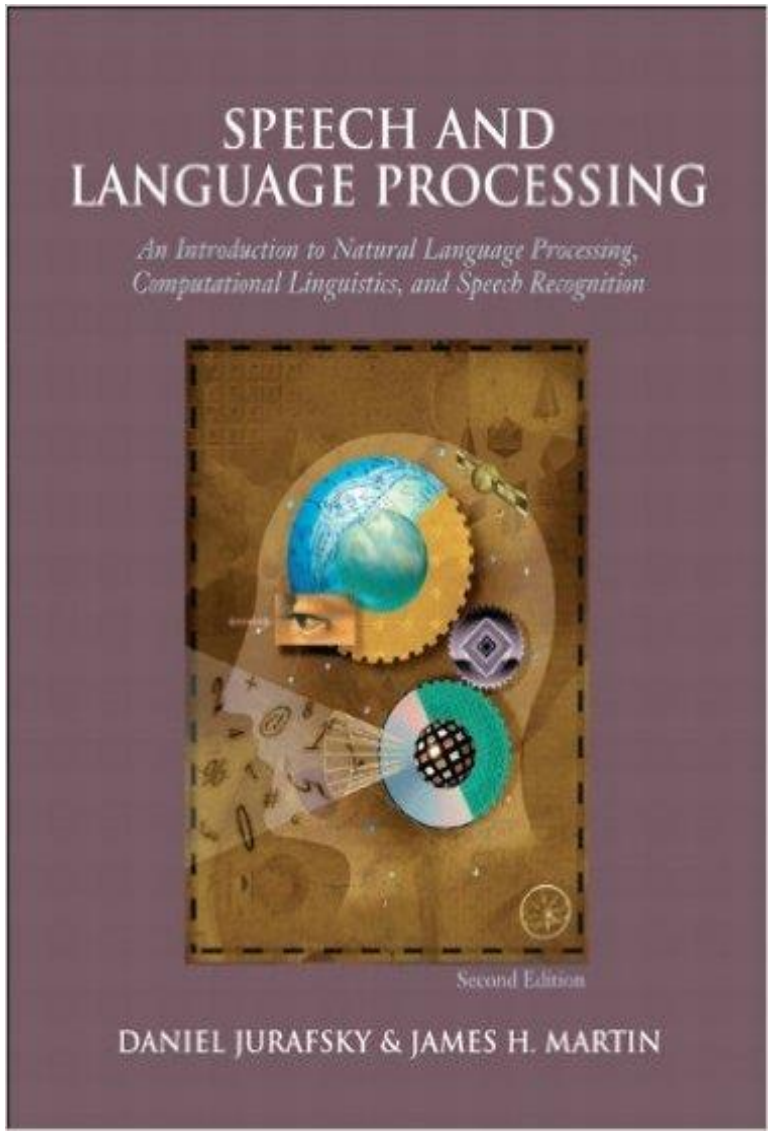


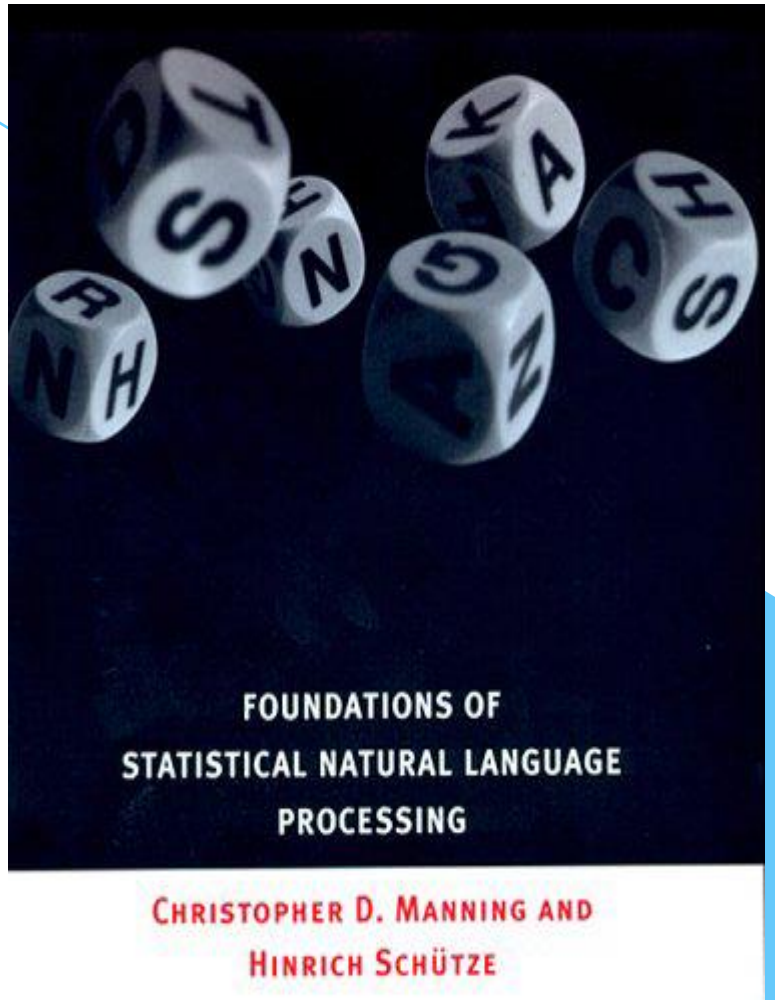


Course content

- 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

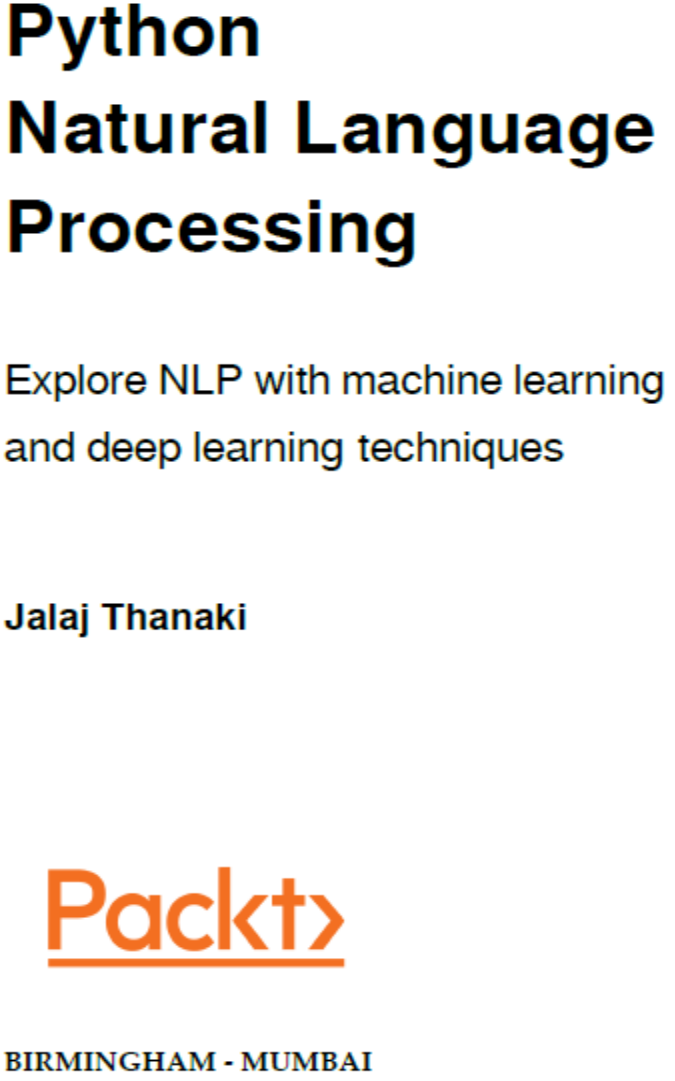
Textbooks (theory)

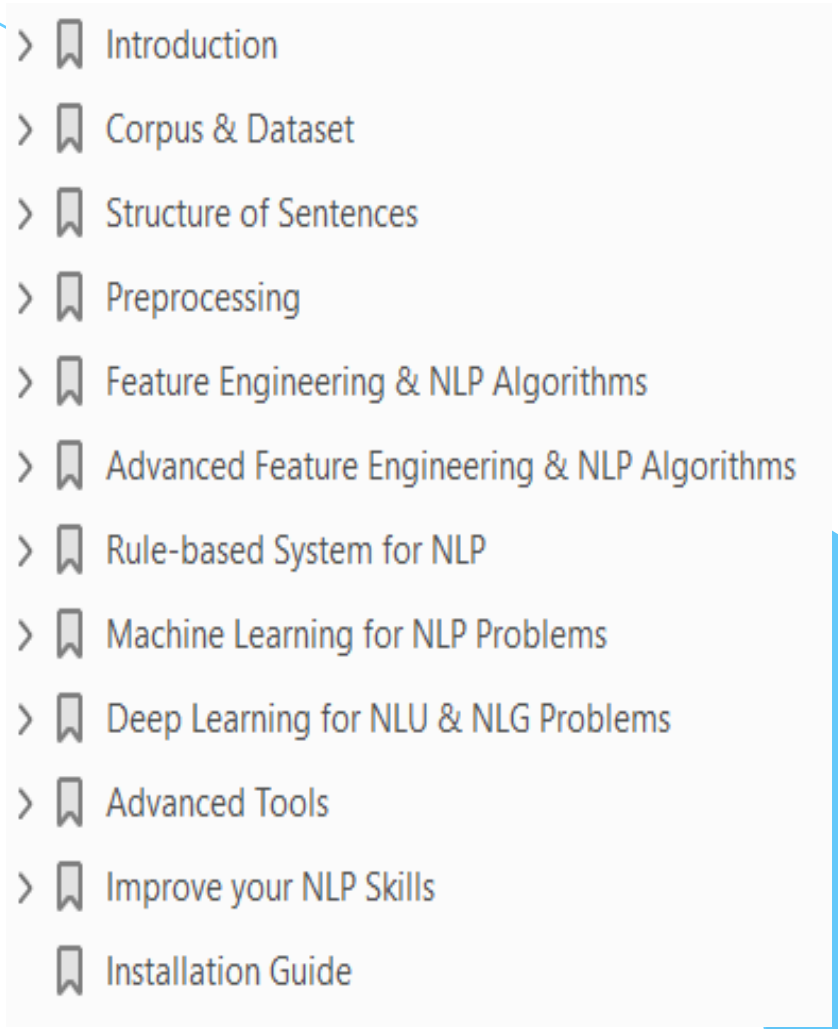




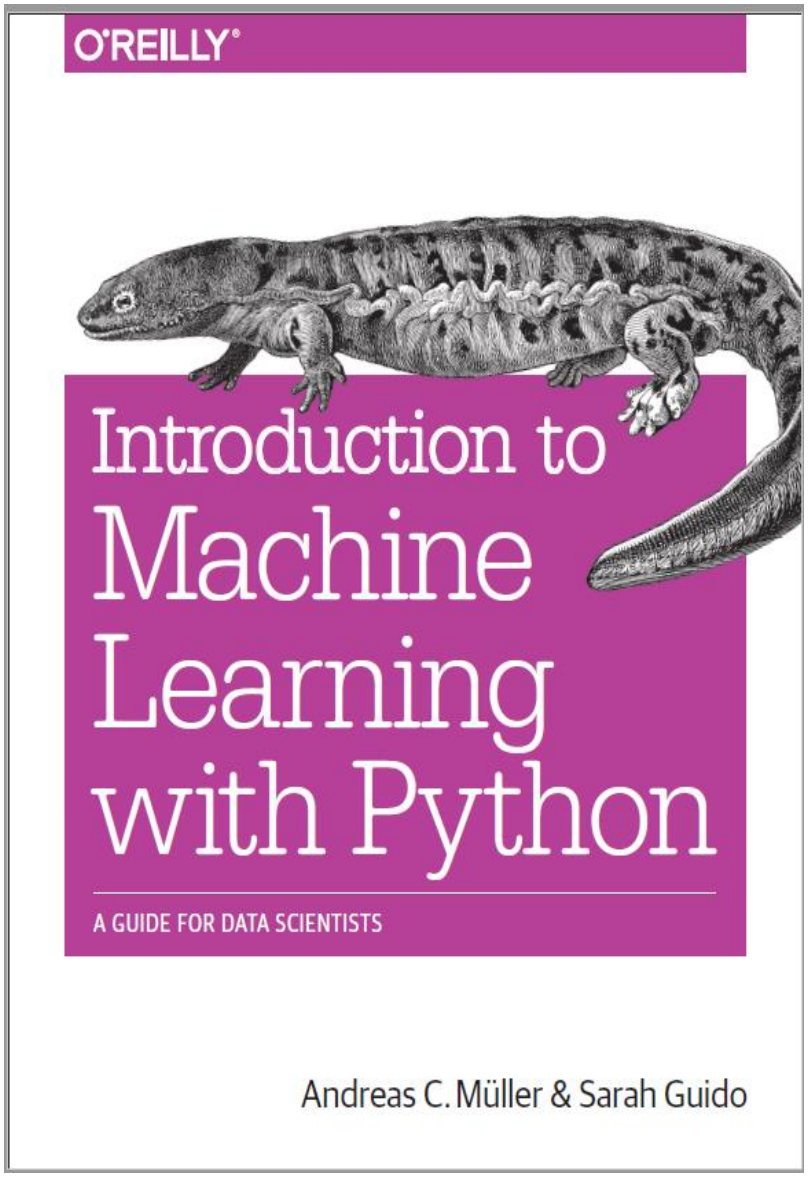
(Coursera textbooks)

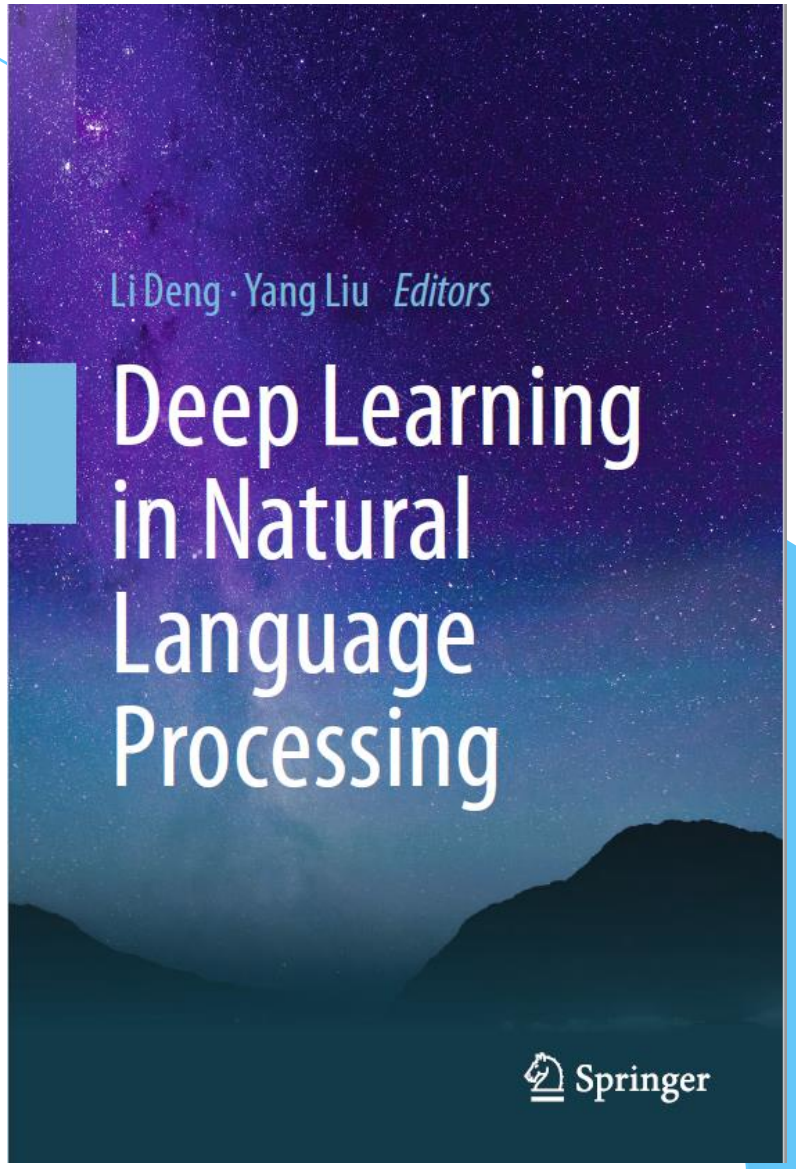
Textbooks (practice)



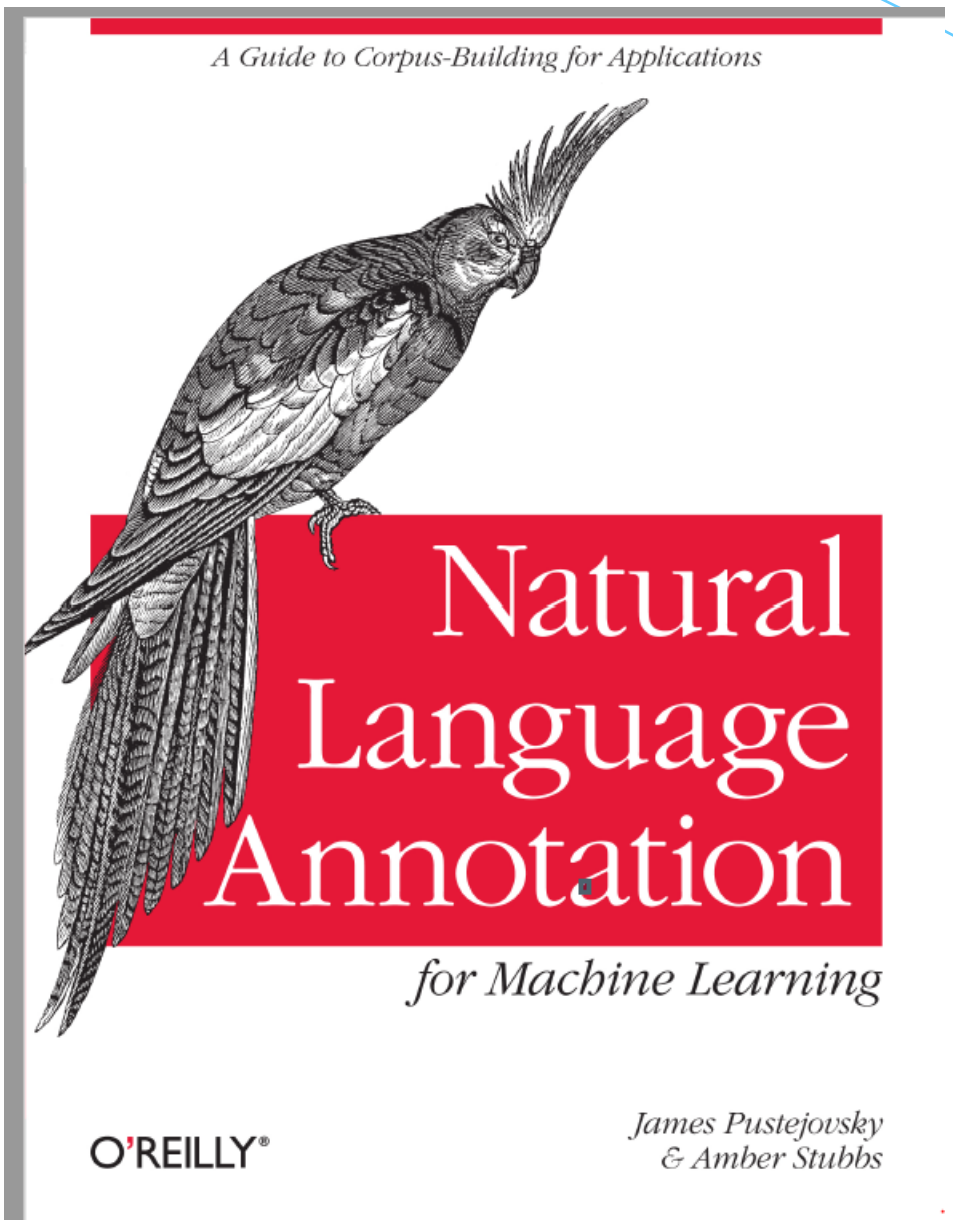


Reference books (ML techniques)



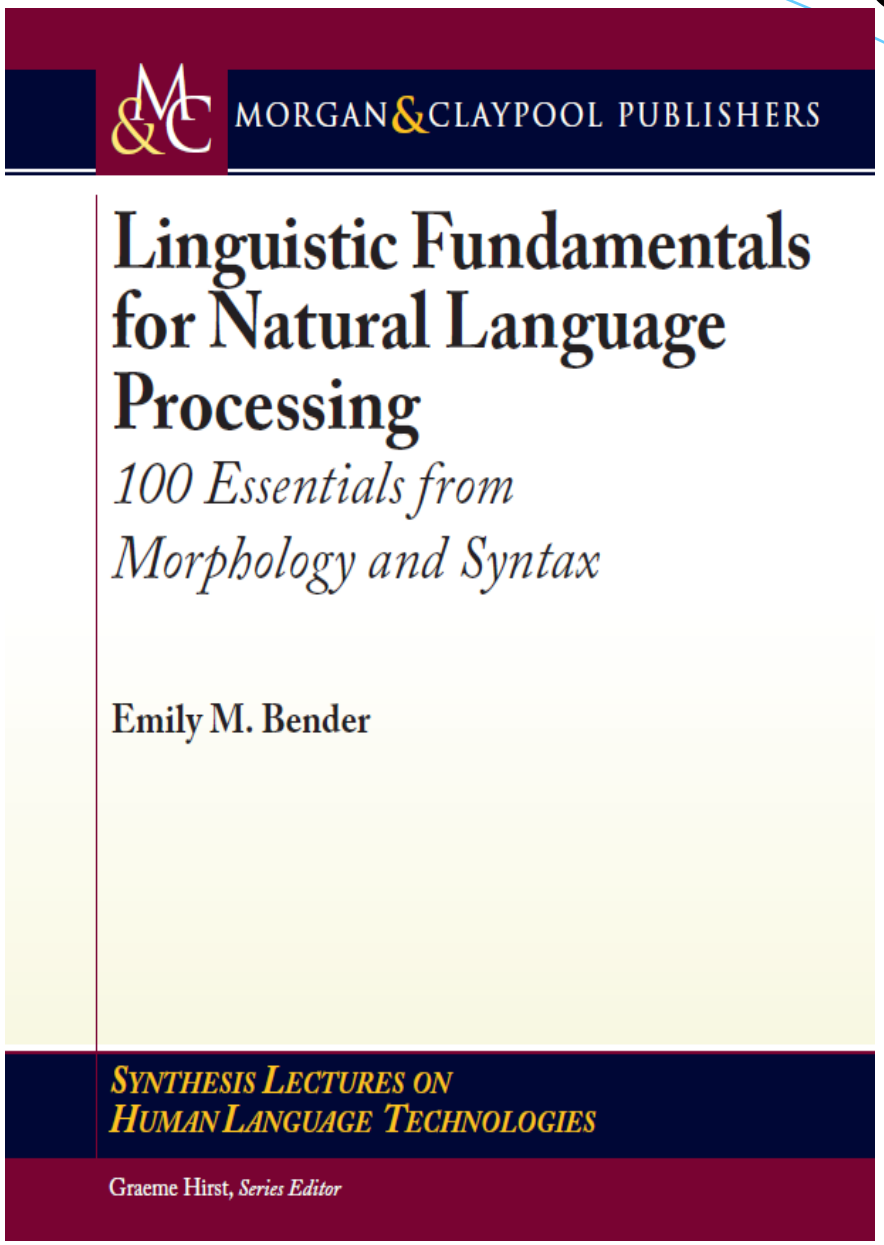


Reference books (training corpus)



- 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

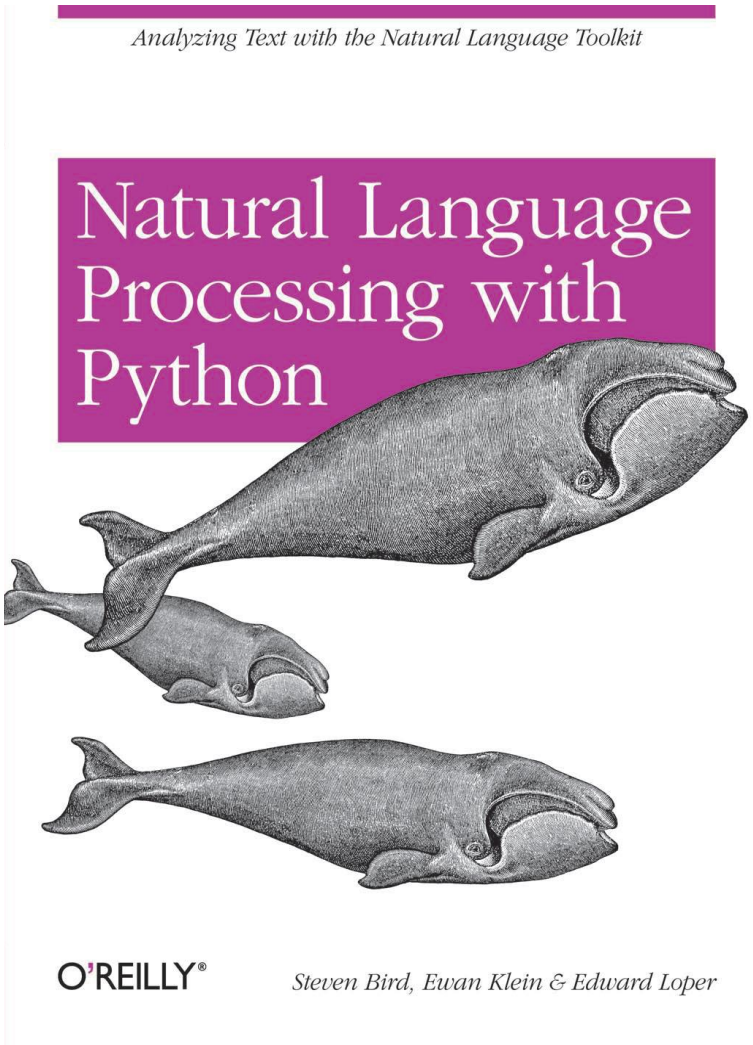
Reference books (linguistics)

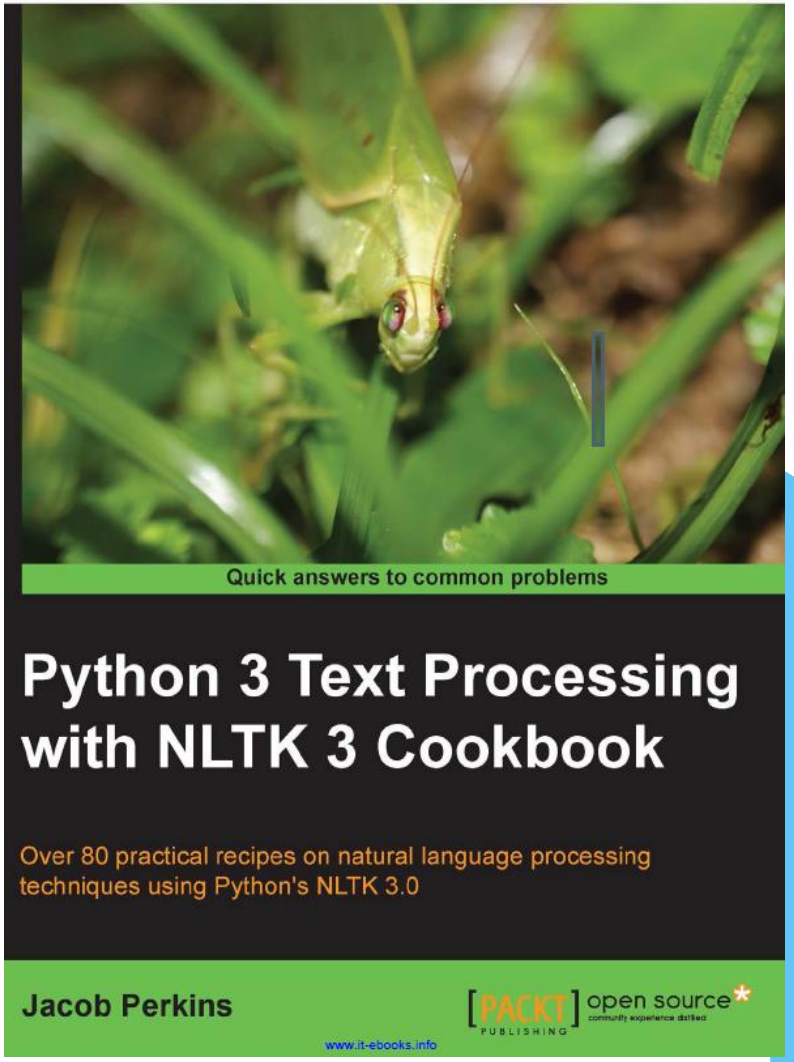


- Introduction/motivation
- Morphology: Introduction
- Morphophonology
- Morphosyntax
- Syntax: Introduction
- Parts of speech
- Heads, arguments and adjuncts
- Argument types and grammatical functions
- Mismatches between syntactic position and semantic roles
- Resources

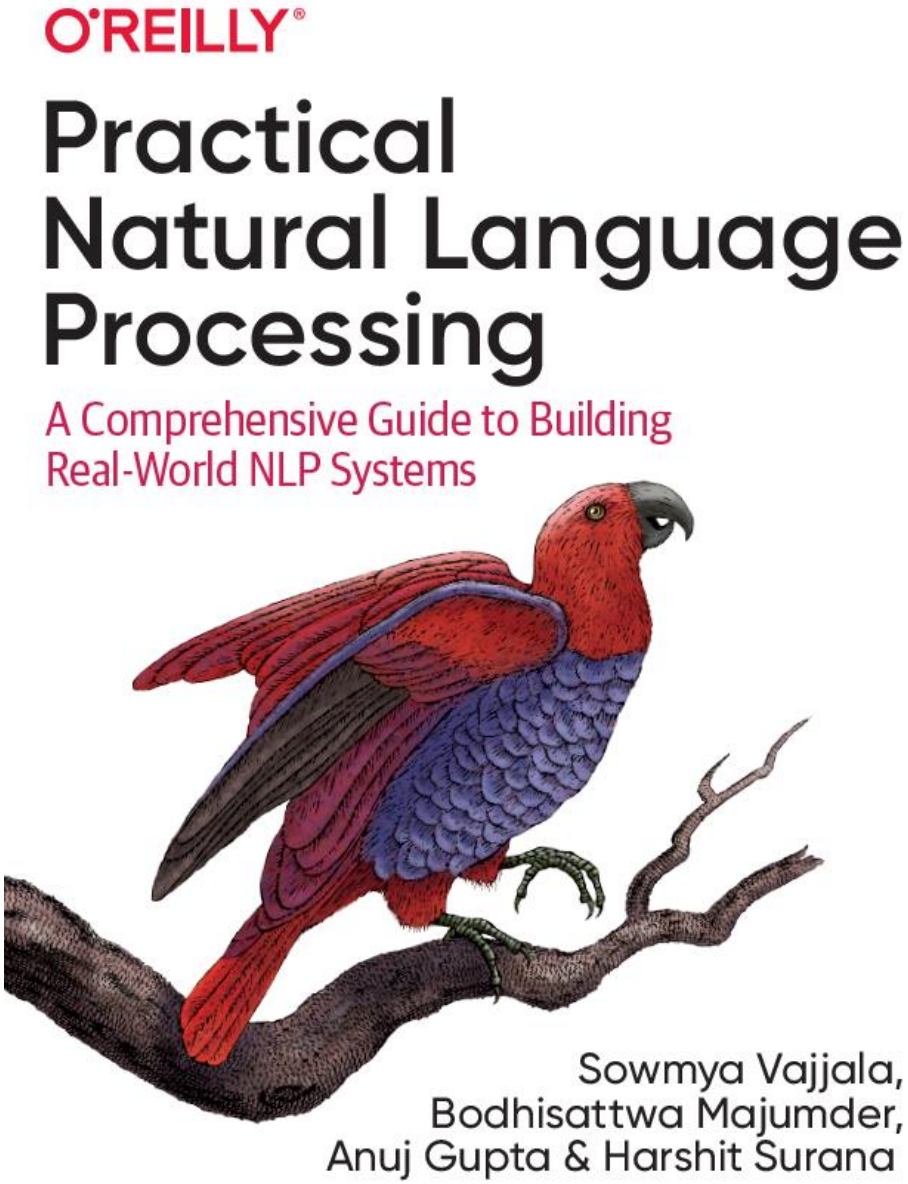
Reference books (Vietnamese)

Reference books (programming)

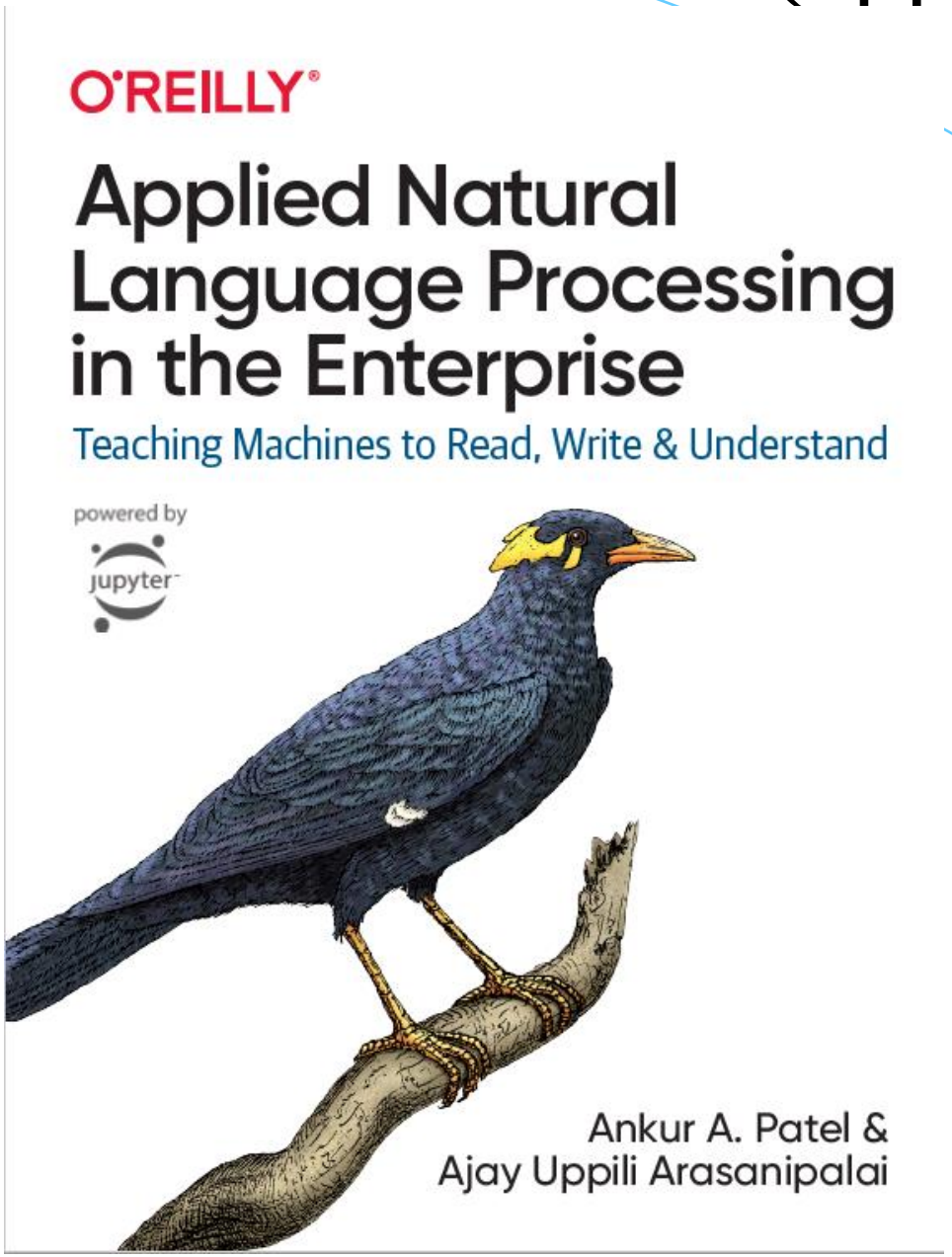




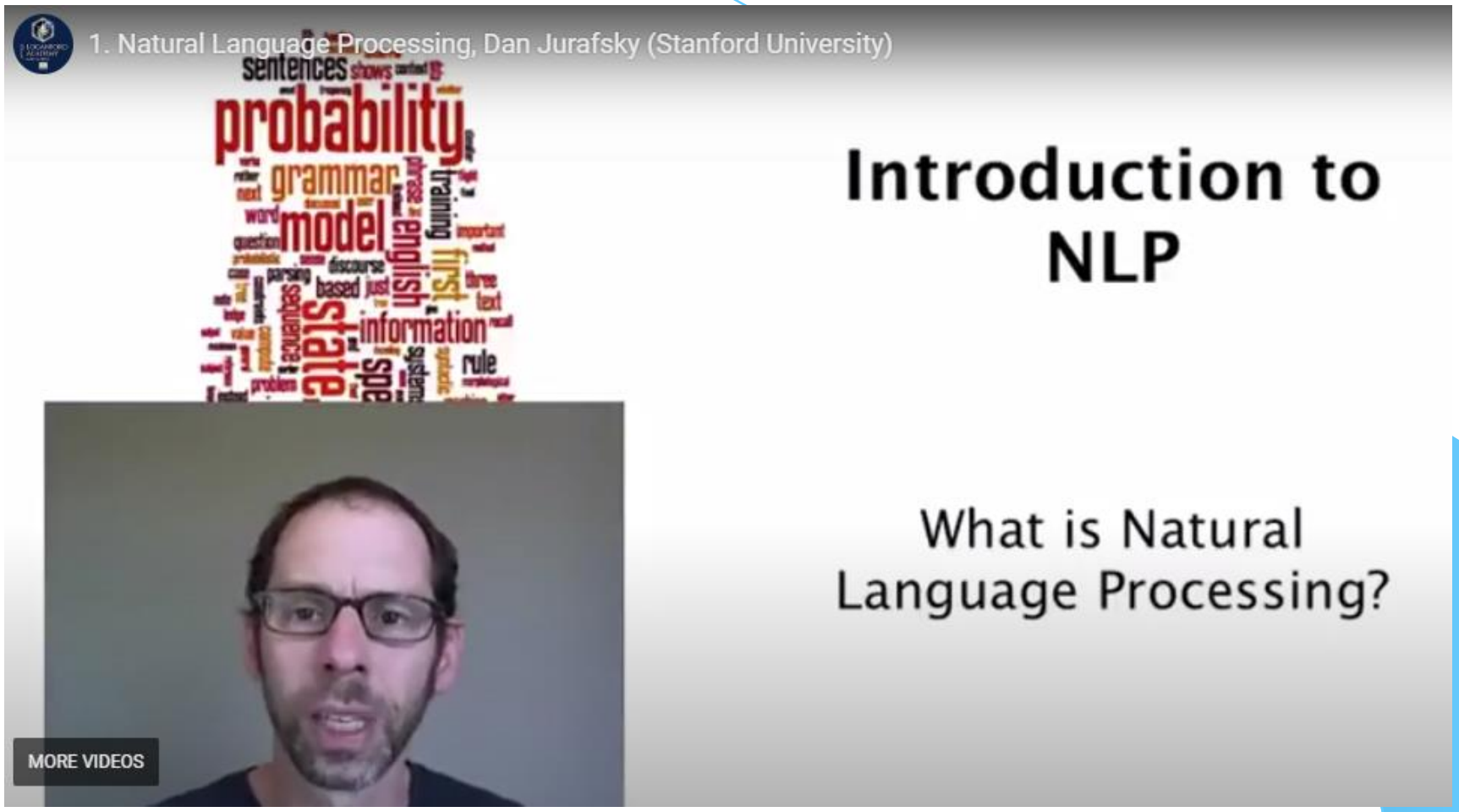
Reference books (applications)



Reference books (applications)



Reference course (Coursera)



➔ 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

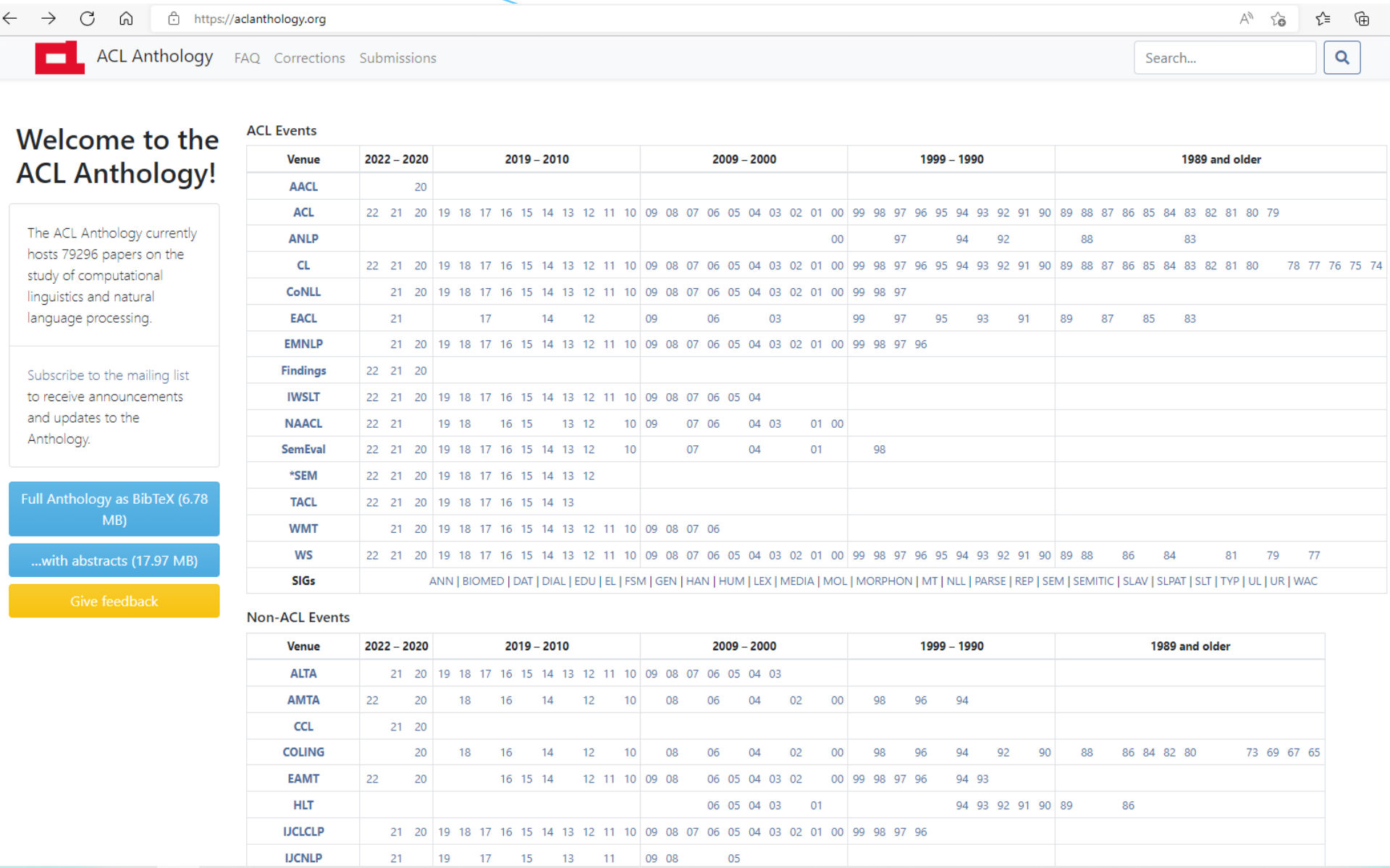
CoNLL: Conference on Computational Natural Language Learning

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).



https://aclanthology.org

ACL Anthology FAQ Corrections Submissions Search...

Welcome to the ACL Anthology!

The ACL Anthology currently hosts 79296 papers on the study of computational linguistics and natural language processing.

Subscribe to the mailing list to receive announcements and updates to the Anthology.

Full Anthology as BibTeX (6.78 MB)

...with abstracts (17.97 MB)

Give feedback

ACL Events

Venue	2022 – 2020	2019 – 2010	2009 – 2000	1999 – 1990	1989 and older
AACL	22 21 20				
ACL	22 21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84 83 82 81 80 79
ANLP			00	97 94 92	88 83
CL	22 21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84 83 82 81 80 78 77 76 75 74
CoNLL	21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97	
EACL	21	17 14 12	09 06 03	99 97 95 93 91	89 87 85 83
EMNLP	21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96	
Findings	22 21 20				
IWSLT	22 21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04		
NAACL	22 21	19 18 16 15 13 12 10	09 07 06 04 03 01 00		
SemEval	22 21 20	19 18 17 16 15 14 13 12 10	07 04 01	98	
*SEM	22 21 20	19 18 17 16 15 14 13 12			
TACL	22 21 20	19 18 17 16 15 14 13			
WMT	21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06		
WS	22 21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 86 84 81 79 77
SIGs	ANN | BIOMED | DAT | DIAL | EDU | EL | FSM | GEN | HAN | HUM | LEX | MEDIA | MOL | MORPHON | MT | NLL | PARSE | REP | SEM | SEMITIC | SLAV | SLPAT | SLT | TYP | UL | UR | WAC				

Non-ACL Events

Venue	2022 – 2020	2019 – 2010	2009 – 2000	1999 – 1990	1989 and older
ALTA	21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03		
AMTA	22 20	18 16 14 12 10	08 06 04 02 00	98 96 94	
CCL	21 20				
COLING	20	18 16 14 12 10	08 06 04 02 00	98 96 94 92 90	88 86 84 82 80 73 69 67 65
EAMT	22 20	16 15 14 12 11 10	09 08 06 05 04 03 02 00	99 98 97 96 94 93	
HLT			06 05 04 03 01	94 93 92 91 90	89 86
IJCLCLP	21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96	
IJCNLP	21	19 17 15 13 11	09 08 05		

Computational Linguistics Journal (2022)

Contents

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 Gül 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 Kálmán | Gábor Gosztolya

pdf bib abs
Deep Learning for Text Style Transfer: A Survey
Di Jin | Zhijing Jin | Zhiting Hu | Olga Vechtomova | Rada Mihalcea