Towards a multi-purpose resource of language corpora

The case of Vietnamese

Hien Pham

University of Alberta

SEALS 23
Chulalongkorn University
Bangkok, Thailand
29 – 31 May, 2013

Outline

Introduction

Methods and Data

Results

- Statistics in corpus linguistics
- Dispersion measures
- Latent Semantic Analysis (LSA)
- Hyperspace Analogue to Language (HAL)

Implications

Conclusion

What is corpus linguistics?

- The study of natural and authentic language data on a large scale
- The computer-aided analysis of extensive collections of texts (transcribed utterances or written texts)
- The main purpose of a corpus is to verify a hypothesis about language
- What corpus linguistics does?
  - provides natural linguistic information
  - facilitates linguistic research
  - provides patterns and collocations of words
  - ...
- What corpus linguistics does not?
  - provide reasons
  - represent the entire language

Corpus linguistics

- Why do we need corpus linguistics?
  - What can linguists do without data?
  - What is the most frequent word in your language?
  - Which are new words in your language?
  - What is the most collocational words with the word “toy”?
  - ...
- Where do word meanings exist?
  - In our brain?
  - In dictionaries?
  - Contexts and collocations: “You shall know a word by the company it keeps” Firth (1957)
  - Quantitative corpus linguistics is used to identify semantic relations of words in contexts

Why do we need other Vietnamese corpora?

- Corpus linguistically, Vietnamese is still an under-studied language
- Constructing corpora is an efficient way of doing language documentation
- Easy to process with corpus tools
- The questions of language acquisition can be revealed with corpus data
- Corpus-based approaches aim at testing and improving theories

Methods and Data

- This study uses (quantitative) corpus linguistic approaches with open-source tools, e.g., Antconc, R, HiDEx, Corsis,... ; and some other programs e.g., WordSmith, SketchEngine,...

- The general corpus
  - Newspaper articles and online short stories
  - 100 million words
  - Tokenized (Lê et al., 2008) and POS-tagged (Lê et al., 2010)
- The film subtitle corpus
  - Movie subtitles
  - 80 million words
  - Tokenized and POS-tagged

A sample of Vietnamese texts

Bởi ông nhận biết được vai trò quan trọng của nhân dân trong việc lưu giữ và bảo tồn những di tích lịch sử của tổ tiên giúp ông hoàn thành được công trình chữ Việt cổ.

— sample.plain.txt

Bởi ông nhận_biết được vai_trò quan_trọng của nhân_dân trong việc lưu_giữ và bảo_tồn những di_tích lịch_sử của tổ_tiên giúp ông hoàn_thành được công_trình chữ Việt cổ .

— sample.tok.txt

Bởi/E ông/N nhận_biết/V được/R vai_trò/N quan_trọng/A của/E nhân_dân/N trong/E việc/N lưu_giữ/V và/CC bảo_tồn/V những/L di_tích/N lịch_sử/A của/E tổ_tiên/N giúp/V ông/N hoàn_thành/V được/R công_trình/N chữ/N Việt/Np cổ/A ./.

— sample.tag.txt

Concordance lines

N	Concordance
1,598	server trước khi đến đích . Một hệ truyền thông điệp truyền tải các yêu cầu và phản hồi thông qua hệ thống dữ liệu . Một thông điệp
1,599	và lãnh thổ của họ không bị ảnh hưởng bởi quy định mới này Các yêu cầu và chi phí visa nhập cảnh vào Mỹ cũng tương tự như ở các
1,600	hạn và tổ chức bộ máy của Bộ Thương mại ; b) Thực hiện các yêu cầu và tạo điều kiện thuận lợi cho cơ quan Nhà nước có thẩm
1,601	sản của mình. Doanh nghiệp bảo hiểm có trách nhiệm thực hiện các yêu cầu và quyết định của Ban kiểm soát khả năng thanh toán. Bộ
1,602	giá về mặt kỹ thuật để chọn danh sách ngắn sẽ được dựa trên các yêu cầu và tiêu chuẩn đánh giá đã quy định trong hồ sơ mời thầu và
1,603	giá về mặt kỹ thuật để chọn danh sách ngắn sẽ được dựa trên các yêu cầu và tiêu chuẩn đánh giá đã quy định trong hồ sơ mời thầu và
1,604	, lời gọi hệ thống cho phép một chương trình đang chạy đưa ra các yêu cầu trực tiếp từ hệ điều hành . Ở mức cơ bản nhất , thông tin
1,605	nó phải có phần thiết kế nền sao cho có thể gia tăng tốc độ theo các yêu cầu trong tương lai . Nó không được tạo ra khí độc hoặc dễ
1,606	phẩm , bưu kiện , báo chí Những công nghệ không đáp ứng các yêu cầu trong các quy định của pháp luật Việt Nam về an toàn lao
1,607	như là một sản phẩm hạ tầng cơ sở nhắm vào một phạm vi rộng các yêu cầu triển khai và phát triển ứng dụng đa cấp . Microsoft
1,608	giao yêu cầu nầy đến Internet sau khi thay đổi địa chỉ IP . Các yêu cầu trên Internet đi qua mạch nầy đến proxy server rồi proxy
1,609	cho quản trị viên . Ngoài ra mrouter còn dùng IGMP để truyền các yêu cầu trên mạng LAN nhằm xác định xem có máy nào còn muốn
1,610	mềm chung trên server , thiết lập cấu hình theo dõi , dự định các yêu cầu trên các máy PC được chỉ định , và thi hành một số tác vụ
1,611	nước cổ phần hoá phải thành lập Hội đồng kiểm kê tài sản theo các yêu cầu trên. Khi một người bị Toà án tuyên bố là đã chết , đã được
1,612	và cơ cấu các loại tiền mặt cần phát hành vào lưu thông theo các yêu cầu trên Căn cứ vào các nội dung quy định tại điểm 2 nêu trên ,
1,613	tutorship xem xét và giới thiệu làm thủ tục xin phép về các yêu cầu trên tại Bộ Văn hoá - Thông tin Uỷ ban nhân dân duy trì việc
1,614	tiếng Việt hoặc tiếng nước ngoài phải làm thủ tục xin phép về các yêu cầu trên tại Bộ Văn hoá - Thông tin 6- Phối hợp với cơ quan , tổ
1,615	phân thành nhiều loại khác nhau : loại điều khiển chương trình , các yêu cầu trạng thái và các yêu cầu nhập / xuất . Đưa ra các dấu ghi
1,616	của người dùng có thể đưa vào cây thư mục . Web server nhận các yêu cầu thông tin từ người sử dụng , sau đó gởi đáp ứng cho họ ,
1,617	dùng chạy các trình duyệt web để nối vào web server và đưa ra các yêu cầu , thông qua nó đến máy tính lớn . Một thủ thuật cũng được
1,618	số kê khai. Môi trường và điều kiện lao động phải đáp ứng các yêu cầu theo quy định của pháp luật. Ngân hàng liên doanh hoạt động
1,619	bộ định tuyến khác để cập nhật bảng định tuyến của nó , trả lời các yêu cầu từ các bộ định tuyến khác , thường xuyên thông báo sự hiện
1,620	giải pháp nầy , việc chạy chương trình trên host tự nó chấp nhận các yêu cầu từ phía PC , và sau đó định dạng lại và gửi chúng đến các

Syntactic pattern

ần). Một thời, giải thưởng ấy không những có giá trị động viên về tinh thần mà còn về vật chất.
điệu, đầy chất thơ, chất Huế, không những làm xao lòng người Huế mà còn làm rung động biết bao c
c đẹp của Trinh. Đối với Bao, không những Trinh có tình yêu mà còn có cả biết ơn và kính phục. S
nh đối với Trần Văn. Đến đây, không những Loan có một công tác nhất định, mà còn được yên tâm vé
của khinh mẹ chồng. Trần Văn không những ghét anh mà còn khinh nữa. Liêm học gạo, đào mỏ, đ
nhỉ? Trần Văn cau mặt. Trinh không những tỏ ra thờ ơ với thời cục mà còn ngu nữa. Hồi yêu Trinh
iệu giục Tú đi tiếp. Quái lạ, không những nó cứ ỳ tại chỗ mà còn trở vai đòn, đứng đối mặt với t
bức xúc nữa là các ngân hàng không những không mua được USD từ kiều hối mà còn phải tốn một lượ
bức xúc nữa là các ngân hàng không những không mua được USD từ kiều hối mà còn phải tốn một lượ
ớc khác. Tư tưởng Hồ Chí Minh không những chinh phục được dân tộc Việt Nam mà còn chinh phục đượ
c ta mỗi ngày một phát triển, không những mạnh ở bên trong mà còn vững ở bên ngoài, được bạn bè
Mỹ lúc đầu còn rụt rè, về sau không những không rụt rè mà còn sáng tác bài hát về Việt Nam, hào
Tin tốt lành này không những là niềm tự hào cho gia đình em mà còn là vinh dự cho c
nh mệnh. Bởi chính từ đây anh không những được giới khoa học hải ngoại biết đến nhiều hơn mà còn
thật bất ngờ. Nhiều buôn làng không những có đời sống kinh tế ổn định mà còn bảo đảm vệ sinh môi
đã có những mâu thuẫn lớn, ta không những không kích cầu mà còn hạn chế cầu vì nếu giờ đưa hết đ

Serial verbs

ng của anh đã tính đến chuyện làm đơn kiến_nghị xin được hoãn di_dời đến sau tết nhưng
_thương nào mới về chợ cũng khổ_sở vì phải hì_hụi gầy dựng tìm mối mới , cả ban quản_lý
Bệnh_viện Bạch_Mai . Tháng sau , tiền hết đành xin về kèm lời khuyên chân_thành của bác
hụ_nữ da đen dắt theo đứa trẻ lũn_cũn tập đi mỉm cười chào tôi . Cảm_giác không an_toàn
ông Dũng kể : " Sinh đứa nào ra cũng tính cho đi học hết ấy chứ . Nhưng túng_thiếu mãi
c thành tài rồi về Nam , nhưng khi đó biết có còn được gặp ba , các em , các anh , các
mệt_lả , đói , lạnh , làm xong là chỉ muốn đổ vật xuống ngủ vùi , có lúc chẳng_thể ăn n
p ông chủ_tịch rồi mà chẳng được gì . Chắc phải đi tiếp lên trung_ương thôi " . Ông khi
e_hơi này thành_công , Như_Mai xin và được giới_thiệu đi học lái máy_bay nhỏ . Cô đã từ
c , nhưng cuối_cùng phải bỏ vì ông_bà thân_sinh sợ phải đi nhận tiền_tuất ! Bà lão gần
òn mệt hơn cả giữ trẻ_con . Ngày_ngày phải cho ăn cho uống vài bận , tắm_rửa chúng hai
m nên giá hơi cao ... ! " . Chúng_tôi yêu_cầu được đưa đi coi mặt vài cô gái nhưng bà C
Tờ_mờ sáng , khát cháy cổ_họng , tôi trở dậy lò_dò đi tìm nước uống . Vén rèm nhìn qua
May_mà năm nay được_mùa chứ nếu không phải đi vay đi mượn , tui nằm đó mà người_ta kéo
ưng lúc đó là 1 g 30 sáng . Tất_cả xe cứu_hỏa được lệnh dãn ra hai bên và chuẩn_bị đèn
chỉ khoảng 6 m 2 , vách bằng ván ép , muốn vào phải đi qua một phòng khác ngăn với phòn
g suốt ngày như_thế . Bởi , ra_vô thì ngại phải đi ngang qua phòng gia_đình người_ta .
một sự lựa_chọn tốt nhất cho học_viên cai_nghiện muốn tiếp_tục theo học bậc đại_học . C
ách_quan : " Địa_phương nơi người sau cai cư_trú phải được xác_nhận là không còn ma_túy
bọn quỉ hiếu_chiến . Vì sao chúng lại thích đi tàn_sát bắn giết những người dân hiền_là
ghé sát mặt để " ngắm " con vật , Dư cầm đưa xa ra kêu : " Chú đừng nhìn gần vì có_thể

Collocations of the word *tim* 'heart'

tim *(-x)* **VietnameseWaC freq = 9396** (72.4 per million)

objectArgument	1835	3.7
suy	165	9.3
đau	287	9.13
mổ	43	8.2
Nhịp	13	7.78
moi	22	7.4
phẫu thuật	19	7.32
van	20	7.3
ghép	23	7.22
muon	10	7.13
thấu	22	7.1
rụng	14	6.67
giải phẫu	9	6.5
Hoà nhịp	5	6.47
thủng	12	6.39
trúng		

simple_sentence_1	1507	3.0
đập	288	9.39
hieu	15	7.9
thổn thức	9	7.25
den	6	6.25
mot	7	6.22
rướm	4	6.22
duoc	7	6.2
mách bảo	4	6.15
cach	4	6.09
ngừng	39	6.02
rung động	6	5.99
rỉ	5	5.97
doc	4	5.92
ứ	4	5.92

modifies_N	269	4.8
nhói	19	10.21
đau nhói	5	8.71
yếu	66	8.52
nhức nhối	6	8.23
buốt	4	7.52
thủng	16	7.38
nát	17	6.98
tái	5	5.7
thấp	11	4.85
thẳng	6	3.7
sâu	4	3.12
gần	7	2.45
đầy	6	2.3
xa	6	2.24

Collocates of the words *tim* 'heart' and *lòng* 'stomach/gut'

modifies_N	269	3045	4.8	5.2
ấm	0	182	0.0	9.8
giàu	0	218	0.0	9.7
cứng	0	77	0.0	8.8
thật	0	349	0.0	8.4
tràn đầy	0	40	0.0	8.1
rộng	0	117	0.0	8.0
quặn	0	25	0.0	8.0
buồn	0	106	0.0	7.8
sâu thẳm	0	20	0.0	7.6
cực	0	36	0.0	7.5
nhẹ	0	61	0.0	7.5
no	0	34	0.0	7.3
dịu	0	19	0.0	7.1
trọn	0	42	0.0	7.1
chung	0	113	0.0	7.0
tan nát	0	14	0.0	7.0
đắng	0	11	0.0	6.7
đầy	6	328	2.3	8.0
nát	17	28	7.0	7.0
yếu	66	68	8.5	8.0
buốt	4	6	7.5	5.8
nhói	19	25	10.2	8.0
thủng	16	0	7.4	0.0
nhức nhối	6	0	8.2	0.0
đau nhói	5	0	8.7	0.0

simple_sentence_1	1507	15898	3.0	3.0
biết ơn	0	672	0.0	10.2
trung thành	0	609	0.0	9.9
ham muốn	0	394	0.0	9.4
tôn kính	0	389	0.0	9.4
thương	0	573	0.0	9.1
kính trọng	0	282	0.0	8.9
trắc ẩn	0	236	0.0	8.9
yêu mến	0	284	0.0	8.9
thương yêu	0	245	0.0	8.7
căm thù	0	199	0.0	8.6
tự hào	0	215	0.0	8.4
từ bi	0	183	0.0	8.3
tự tin	0	187	0.0	8.3
nhân ái	0	153	0.0	8.2
tin tưởng	0	227	0.0	8.2
tự ái	0	137	0.0	8.0
tham	0	188	0.0	7.8
khao khát	0	105	0.0	7.5
bi	0	119	0.0	7.4
tin	0	366	0.0	7.3
quyết tâm	0	117	0.0	7.3
yêu	20	1730	3.8	9.8
yêu thương	16	437	5.6	9.1
hiếu	15	0	7.9	0.0
đập	288	0	9.4	0.0

Syllable frequency and dispersion

- Quantitative corpus linguistics is characterized by the exhaustive and systematic analysis of linguistic phenomena on the basis of a linguistic corpus.
- Many types of frequencies and dispersion measures can be computed.

Table: Frequencies and dispersion measures for Vietnamese syllables.

Syllabeme	SyllFreq	SyllDisp	SyllDisp%	SyllPerMil	LogSyllFreq	LogSyllDisp
thuỷ	13177	4001	29.91	166.80	4.12	3.60
thuỵ	1100	618	4.62	13.92	3.04	2.79
thuyên	676	433	3.24	8.56	2.83	2.64
thuyền	17282	3492	26.10	218.76	4.24	3.54
thuyết	12346	5699	42.60	156.28	4.09	3.76
ti	1629	1056	7.89	20.62	3.21	3.02

Word frequency and dispersion

Table: Frequencies and dispersion measures for Vietnamese words. WF: WordFreq, WD: WordDisp WDP: WordDispPercent, LSF: LogSubtFreq, LSD: LogSubtDisp, WPM: WFPerMillion

Word	WF	WD	WDP	LSF	LSD	WPM
á à	87	82	0.61	1.94	1.92	1.10
gian ác	46	38	0.28	1.67	1.59	0.58
tội ác	2422	1568	11.57	3.38	3.20	30.66
hiểm ác	64	61	0.45	1.81	1.79	0.81
bạc ác	2	2	0.01	0.48	0.48	0.03
quái ác	40	40	0.30	1.61	1.61	0.51

Dispersion measures (Gries, 2009)

Abbreviation	Measure
FREQ	observed frequency of word w
RANGE	number of parts with word w
MAXMIN	max. freq. of w/part—min. freq. of w/part
SD	standard deviation of frequencies
VARCOEFF	variation coefficient of frequencies
CHISQUARE	chi-square value of the frequency distribution
D_EQ	Juilland et al.'s D (assuming equal parts)
D_UNEQ	Juilland et al.'s D (not assuming equal parts)
D2	Carroll's D_2
S_EQ	Rosengren's S (assuming equal parts)
S_UNEQ	Rosengren's S (not assuming equal parts)
D3	Lyne's D_3
DC	Distributional Consistency
IDF	Inverse Document Frequency
ENGVALL	Engvall's measure
U_EQ	Juilland et al.'s usage coefficient U (assuming equal parts)
U_UNEQ	Juilland et al.'s usage coefficient U (not assuming equal parts)
UM_CARR	Carroll's U_m
AF_EQ	Rosengren's Adjusted Frequency AF (assuming equal parts)
AF_UNEQ	Rosengren's Adjusted Frequency AF (not assuming equal
parts) Ur_KROM	Kromer's U_R
F_ARF	Savický and Hlaváčová's f_{ARF}
AWT	Savický and Hlaváčová's AWT
F_AWT	Savický and Hlaváčová's f_{AWT}
ALD	Savický and Hlaváčová's ALD
F_ALD	Savický and Hlaváčová's f_{ALD}
SELF_DISP	Washtell's self-dispersion
D	

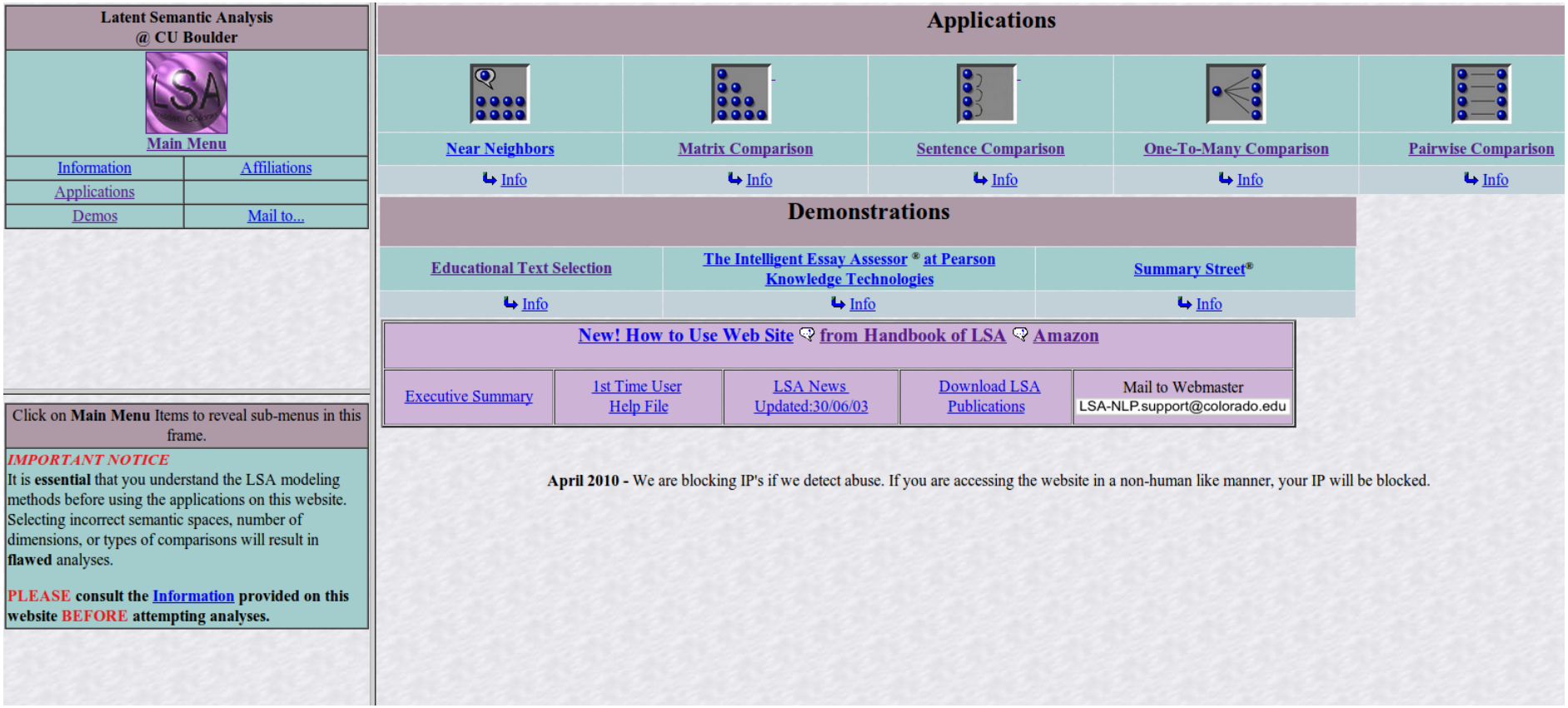
Dispersion measures

Word	FREQ	RANGE	MAXMIN	SD	VARCOEFF	CHISQUARE	D_EQ	D_UNEQ
cảnh gần	8.00	8.00	1.00	0.01	147.34	220704.86	0.65	0.53
cánh phấn	2.00	2.00	1.00	0.00	294.69	15572.46	0.29	0.23
cảnh sát	25695.00	12434.00	46.00	0.76	5.16	869451.64	0.99	0.99
cao lương	100.00	49.00	23.00	0.08	135.55	575687.79	0.67	0.75
cặp lồng	30.00	23.00	3.00	0.02	94.21	191176.18	0.77	0.64
cạp nia	9.00	7.00	3.00	0.01	179.34	137331.11	0.57	0.51

D2	S_EQ	S_UNEQ	D3	DC	IDF	ENGVALL	U_EQ	U_UNEQ	UM_CARR
0.17	0.00	0.00	-5426.44	0.00	14.41	0.00	5.17	4.28	1.38
0.06	0.00	0.00	-21709.50	0.00	16.41	0.00	0.59	0.46	0.11
0.76	0.06	0.05	-5.66	0.06	3.80	1839.48	25376.75	25394.28	19446.87
0.26	0.00	0.00	-4592.74	0.00	11.79	0.03	67.47	74.61	25.61
0.25	0.00	0.00	-2218.07	0.00	12.88	0.00	23.22	19.08	7.61
0.15	0.00	0.00	-8039.77	0.00	14.60	0.00	5.13	4.63	1.37

AF_EQ	AF_UNEQ	Ur_KROM	F_ARF	AWT	F_AWT	ALD	F_ALD	DP	DPnorm
0.00	0.00	8.00	5.63	7427879.64	5.72	7.13	6.27	1.00	1.00
0.00	0.00	2.00	1.20	34989851.57	1.21	7.79	1.38	1.00	1.00
1606.13	1393.76	17028.34	9384.85	20117.13	2111.81	4.15	5980.62	0.93	0.93
0.02	0.08	58.00	34.54	2075863.59	20.46	6.49	27.38	1.00	1.00
0.00	0.02	26.33	15.24	3737474.52	11.37	6.77	14.35	1.00	1.00
0.00	0.00	7.83	4.32	11980483.74	3.55	7.30	4.29	1.00	1.00

Latent Semantic Analysis: English (Landauer and Dumais, 1997)

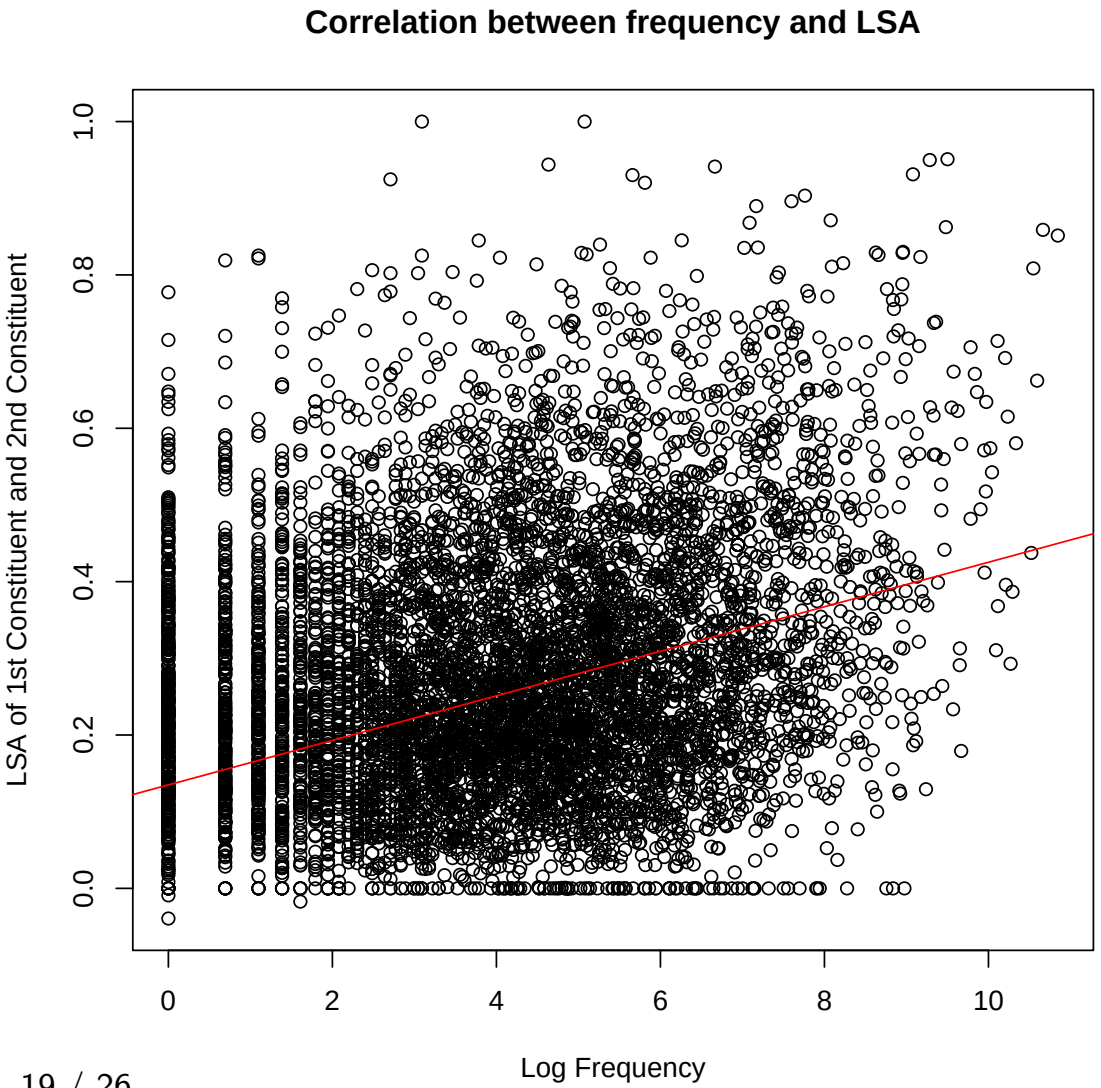


http://lsa.colorado.edu/

Latent Semantic Analysis (LSA)

- LSA is one of a growing number of corpus-based techniques that employ statistical machine learning in text analysis
- Semantic content of a document can be usefully approximated as a sum of the meaning of its words
- How does LSA work?
  - Documents are represented as "bags of words", order of words in a document is not important, only how many times each word appears in a document
  - Concepts are represented as patterns of words that usually appear together in documents.
  - Words are assumed to have only one meaning. This is clearly not the case (e.g., the word *banks* could be *river banks* or *financial banks*) but it makes the problem tractable.

Latent Semantic Analysis



Hyperspace Analogue to Language (HAL): English (Burgess and Lund, 1997)

`http://locutus.ucr.edu/hds.html`

Hyperspace Analogue to Language (HAL)

- HAL is also a type of semantic memory, developed by Kevin Lund and Curt Burgess in 1996.
- HAL considers context only as the words that immediately surround a given word. HAL computes an NxN matrix, where N is the number of words in its lexicon, using a 10-word reading frame that moves incrementally through a corpus of text.
- Two words are simultaneously in the frame, the association between them is increased, that is, the corresponding cell in the NxN matrix is incremented.
- As in LSA, the semantic similarity between two words is given by the cosine of the angle between their vectors (dimension reduction may be performed on this matrix, as well).
- In HAL, then, two words are semantically related if they tend to appear with the same words.

Implications

- The parts of the corpora can be used in compiling dictionaries, reference grammars, teaching and learning Vietnamese, among other linguistic research fields.
  - The linguistic information extracted from corpora can be used in designing curriculum and in language teaching and learning.
  - The frequency lists can be used as a source for compiling language textbook.
  - Language documentation
  - ...
- The statistical-semantic measures, such as *frequencies, dispersion measures, LSA scores and HAL scores* can be used in psycholinguistics and studying human memory.
  - Automated essay scoring
  - Cognitive linguistics, Psycholinguistics, Computer Science

Conclusion

- The availability of data and the current computational power allow us to construct complex and multi-facet linguistic corpora and/or linguistic databases.
- An interdisciplinary approach reveals insight into the nature of language acquisition and language processing.
- Thanks to the corpus tools, linguistic data can be mined and used efficiently.
- The outcomes of quantitative corpus linguisitcs are food for quantitative linguistics, psycholinguistics, cognitive linguistics, etc.
- ...

References (1)

Burgess, C. and Lund, K. (1997). Modelling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, 12(2-3):177–210.

Firth, J. R. (1957). *Papers in linguistics, 1934-1951*. Oxford University Press, London.

Gries, S. T. (2009). Dispersions and adjusted frequencies in corpora: Further explorations. *Language and Computers*, 71(1):197–212.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

Lê, H. P., Nguyen, T. M. H., Roussanaly, A., and Ho, V. (2008). A hybrid approach to word segmentation of Vietnamese texts. In Martin-Vide, C., Otto, F., and Fernau, H., editors, *Language and automata theory and applications*, volume 5196 of *Lecture Notes in Computer Science*, pages 240–249. Springer Berlin, Heidelberg.

References (2)

Lê, H. P., Roussanaly, A., Nguyen, T. M. H., and Rossignol, M. (2010). An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts. In *Traitement Automatique des Langues Naturelles - TALN 2010*, page 12, Montréal Canada. ATALA (Association pour le Traitement Automatique des Langues).

Thank you!