

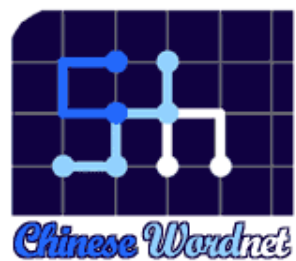
Transliteration Variants: Corpus-based Studies and Sociolinguistic Observations

外來詞音譯對比：
語料庫為本的研究與社會語言學的觀察

Chu-Ren Huang 黃居仁

Academia Sinica

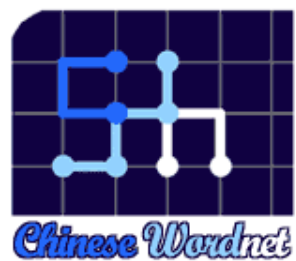
<http://cwn.ling.sinica.edu.tw/huang/huang.htm>



Outline

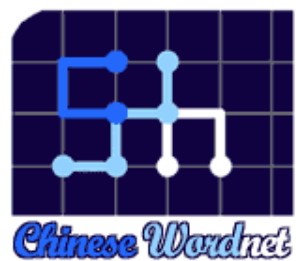
- Observation and Motivation
- Introducing *Comparable Corpora*
 - And the *Chinese Gigaword Corpus*
 - With *Chinese WordSketch*
- Automatic Discovery of Transliteration Variants from Three Chinese Speaking Societies
- Preliminary Analyses
- Concluding Remarks

語音



Observation: transliterated names vary

- **Clinton can be transliterated differently in different Chinese speaking societies**
- 克林頓 vs. 柯林頓
- **And in each society, its top five most salient partner in the and/or constructions are**



What can transliteration variations tell us?

For 克林頓

- 戈爾/Gore 萊溫斯基/Lewinsky
- 布什/Bush 葉利欽/ Yeltin

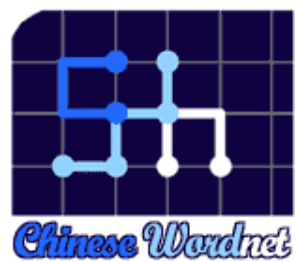
For 柯林頓

- 布希 Bush 葉爾勤 Yeltin
- 高爾 Gore 呂茵斯基 /呂女 Lewinsky

語音

A Challenge

- 克林頓/克林敦/柯林頓/柯林斯/柯琳頓/...
- Lewinsky in the news (Taiwan): 呂茵斯基、呂文絲基、呂茵斯、陸文斯基, 陸茵斯基、柳思基、陸雯絲姬、陸文斯基、呂茵斯基、露文斯基、李文斯基、露溫斯基、羅恩斯基、李雯斯基. .
- **How to judge which two forms are the same (or different)? 異中如何求同?**
- **How to compare salient collocations of different forms?**
- **Large comparable Corpora to provide context and supporting evidence for contrast**



Comparable Corpus 對應語料庫

- **A comparable corpus is one which selects similar texts in more than one language or variety.**
 - **The possibilities of a comparable corpus are to compare different languages or varieties in similar circumstances of communication, but avoiding the inevitable distortion introduced by the translations of a parallel corpus.**
 - **Ex. International Corpus of English (ICE), LiVAC**



A Comparable Corpus for Varieties of Chinese

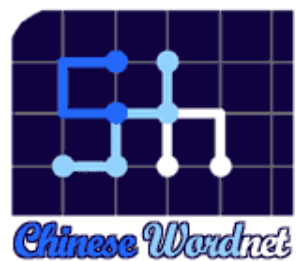
Chinese Gigaword Second Edition (2005)

- Produced and released by Linguistic Data Consortium (LDC) in 2003 (first edition).
- Newswire text data in Chinese.
- Three distinct international sources :
 - Central News Agency of Taiwan
 - Xinhua News Agency of Beijing
 - Zaobao Newspaper of Singapore
- Covering the same years and months



Coverage of Chinese GigaWord Corpus

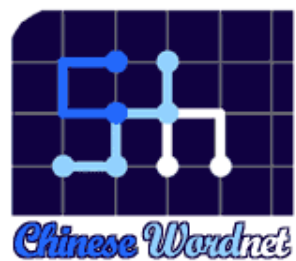
	CNA	Xinhua	Zaobao
First Edition	1991- 2002	1990- 2002	
New in Second Edition	Oct. 2002 - Dec. 2004	Jan. 2003 - Dec. 2004	Oct. 2000 - Sep. 2003



CGW Corpus Data Format

All text data are presented in SGML form, using a very simple, minimal markup structure.

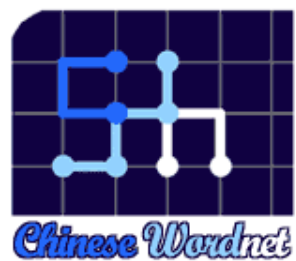
```
<DOC id="CNA19910101.0003" type="story">
<HEADLINE>
捷運局對工程噪音採多項防治措施
</HEADLINE>
<DATELINE>
(中央社台北一日電)
</DATELINE>
<TEXT>
<P>
台北都會區捷運工程正處於積極趕工階段,...
</P>
<P>
淡水線工程進度百分之三十六點一九,落後百分之二點六七,...
</P>
</TEXT>
</DOC>
```



Basic Statistics of CGW Corpus

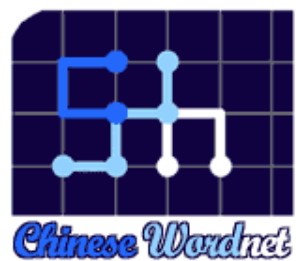
	Resource	Characters	Words	Documents
First Edition	CNA	735	462	1,649
	Xinhua	382	252	817
	TOTAL	1,118	714	2,466
Second Edition	CNA	792	497	1,769
	Xinhua	471	310	992
	Zaobao	28	18	41
	TOTAL	1,291	825	2,803

Unit for character/word : **Million**



CGW Word Types/Tokens after automatic tagging by AS

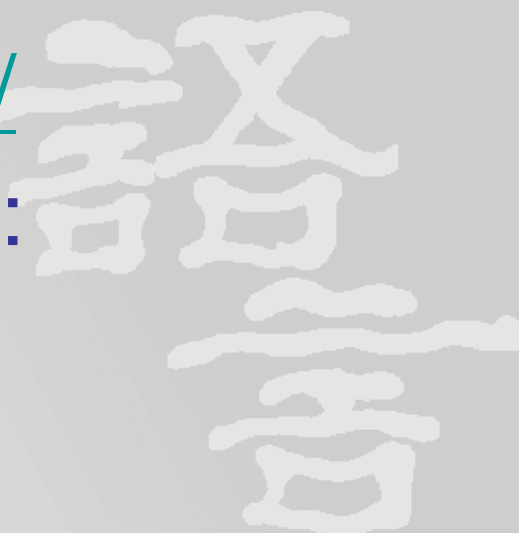
	Word Type	Word Token
CNA	1,917,093	496,465,879
XIN	1,409,747	305,595,420
ZBN	273,111	18,328,571
Total	2,999,590	820,389,870



Chinese WordSketch

<http://wordsketch.ling.sinica.edu.tw/>

- Chinese GigaWord Corpus 2.0:
1,400,000,000 characters
1990-2004 (LDC, 2005)
- Sinica Corpus 5.0
 - 10,000,000 words
- In collaboration with the SketchEngine team lead by Adam Kilgarriff's team at Lexical Computing (www.sketchengine.co.uk)



中文詞彙特性速描系統簡介

中文詞彙特性速描系統是一個結合了鉅量語料庫的語法知識產生系統。

在中文詞彙特性速描系統上除了一般的關鍵詞及語境查詢外，更提供了詞彙特性速描(word sketches)、語法關係以及同近義詞分析等自動產生的語法知識。「中文詞彙特性速描系統」與十四億字的LDC Chinese Gigaword語料庫結合後，提供了絕大部分中文詞彙實際使用的規則性描述，可應用於辭典編撰、華語文教學、語言學研究與自然語言處理。

最新消息

- 中文詞彙特性速描系統開放國內相關研究人員申請使用，請下載申請表格填寫後寄至 cwn@gate.sinica.edu.tw

線上系統

- 中央研究院中文詞彙速描系統
<http://wordsketch.ling.sinica.edu.tw/>
- Word Sketch Engine
<http://www.sketchengine.co.uk/>

工作小組主要人員

黃居仁	中央研究院語言學研究所研究員
Simon Smith	銘傳大學助理教授
馬偉雲	Columbia University, graduate student
Petr Šimon	TIGP, Academia Sinica, Ph.D. student

語料庫

- 中文十億詞語料庫(Chinese GigaWordCorpus)
- 中央研究院平衡語料庫5.0版

相關文件

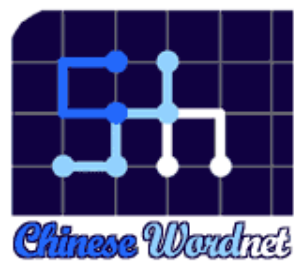
中文詞彙特性速描系統由中央研究院語言學研究所 [中文詞彙網路小組](#) 開發管理

THE SKETCH ENGINE IS PROVIDED BY LEXICAL COMPUTING LTD
THE GIGAWORD IS PROVIDED BY LINGUISTIC DATA CONSORTIUM
詞類標記版權屬中央研究院所有
中文詞彙特性速描系統語法規則版權屬中央研究院所有

[申請表格下載](#)




[常見問題](#)

[問題反應與意見交流](#)



CWS: Concordance

[Home](#) [Concordance](#) [Word Sketch](#) [Thesaurus](#) [Sketch-Diff](#) [Frequency](#) [Collocation](#)

[KWIC/Sentence](#) [View options](#) [Sample](#) [Filter](#) [Sort](#)   

Page of 1003 [Go](#)

[Next](#) | [Last](#)

Corpus: [gigaw](#)
Hits: 20057
[conc description](#)

[CNA CMN 19910103.0268](#)

[CNA CMN 19910104.0079](#)

[CNA CMN 19910105.0141](#)

[CNA CMN 19910108.0021](#)

[CNA CMN 19910111.0195](#)

[CNA CMN 19910116.0003](#)

[CNA CMN 19910117.0239](#)

[CNA CMN 19910118.0154](#)

[CNA CMN 19910121.0029](#)

[CNA CMN 19910121.0066](#)

[CNA CMN 19910123.0131](#)

[CNA CMN 19910123.0278](#)

[CNA CMN 19910124.0271](#)

[CNA CMN 19910126.0240](#)

[CNA CMN 19910128.0235](#)

[CNA CMN 19910129.0021](#)

，南韓與阿拉伯聯合大公國連一分都未得到。埃及與喀麥隆，走遍了世界各地，甚至連居家都遷徙不定，有時在，未來不但老人人口增加，連小孩人口都增加，依賴人口二十來人擠於一室，睡覺時連翻身的餘地都沒有。夏天向錢看，奢靡之風過度，連大陸同胞都看不起我們。在該校就讀，三年前該校連一位中國留學生都沒有。中子群，它的穿透力非常強，連坦克車的鋼板都擋不住，閱報欄和公車上議論紛紛。連一些老年人都在探詢，美軍將會沒水、沒電、實際上連衛生設備都沒有，汽油也(伊拉克)還剩下多少，甚至連開始時擁有多少都有不同此事一無所悉。他也相信連總經理都不清楚的事，其它伊拉克的巴斯拉港，強烈爆炸聲連鄰國伊朗都能聽得到。</p><p>猛烈的空、海轟炸，強度連鄰近的伊朗城市都可以感受到轟炸，精確的程度，連伊拉克軍都感到震驚。</p><p>一波波轟炸，形成的漫天煙柱連鄰近的伊朗省分都可以看懷疑時，沙丹·胡笙說，「連百萬分之一的懷疑都沒有。」



Design Criteria of Sketch Engine

- The Word Sketch Engine takes as input a corpus of any language and a corresponding grammar patterns and generates word sketches for the words of that language
- Ranked by Saliency: frequency adjusted MI
- Automatic discovery of grammatical relations is hard, but there is a solution when corpus is big enough



From Lexical Types to Relations Types

- BNC has 100 million Words
 - 939,028 word types
 - 70,000,000 tuples (relations) Extracted
 - More than 70 relations per lemma
- For CWS II, and CGW corpus (nearly 500 million words in CNA data)
 - 1,917,093 word Types
 - 59,183,238 tuples (<eat, obj, rice>)
 - More than 30 relations per lemma

語音

Home Concordance **Word Sketch** Thesaurus Sketch-Diff

發票 chinese_all_trd:taiwan-only freq = 8408 [change options](#)

object_of 928 5.2		subject_of 692 0.3		a_modifier 276 0.5		n_modifier 562 -12.3		modifies 607 -12.8	
開立	253 73.78	收執	15 43.41	不實	85 57.41	增值稅	103 47.76	逃漏稅	16 30.08
虛開	50 51.4	給	69 32.61	假	48 46.83	銷貨	17 34.96	存根聯	7 29.53
偽造	70 40.88	對獎	7 29.59	空白	24 41.9	收銀機	16 34.19	金額	55 27.38
虛購	12 40.47	中獎	12 25.43	中獎	22 39.32	式	32 30.07	面額	13 25.7
漏開	16 40.07	逃漏	9 24.03	普通	19 28.32	小額	23 29.64	日期	18 21.74
開具	30 39.48	充當	9 23.46	領用	5 27.62	進項	9 27.8	獎金	18 20.77
變造	14 27.67	兌換	16 23.27	可疑	5 15.34	聯	29 27.75	偷稅	7 20.53
虛設	7 22.83	換版工作	3 22.92	增值	3 12.82	虛立	4 26.96	助創世	2 19.68
使用	49 22.27	捐贈	13 22.15	原始	3 11.24	加副聯	3 22.14	號碼	12 19.14
開出	15 21.51	換好	4 21.78	填開式	1 11.17	愛心	18 21.81	影本	7 18.24
購買	24 20.21	抬頭	7 20.97	免用	1 10.42	盜竊	11 21.63	案件	22 17.59
取得	34 19.33	犯罪	21 20.69	全額	2 10.38	票載	3 20.76	憑證	8 17.17
募集	9 17.95	膨脹	7 19.8	小小	2 10.3	六獎	3 20.36	人因	8 17.02
印製	7 16.2	傳情	4 19.13	正規	2 10.2	開假	3 20.36	管理員	7 16.76
持	13 15.4	冒領	5 18.31	原	5 9.97	電腦版	3 19.02	收據	5 15.47
假造	3 14.93	盜領	5 18.16	欣榮	1 8.77	票券	11 17.95	魔方	2 15.45
發出去	2 14.13	捐給	5 17.81	作廢	1 7.66	開具假	2 17.84	普獎	2 14.52
買	9 13.66	開立	6 15.83	足額	1 7.0	預前	2 17.84	婚紗秀	2 14.07



Home Concordance Word Sketch Thesaurus **Sketch-Diff**

Word Sketch Differences Entry Form

Corpus: chinese_all_trd

First lemma: 明星

Second lemma: 演員

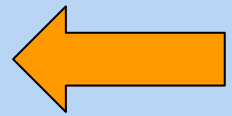
Sort grammatical relations:

Separate blocks: all in one block common/exclusive blocks

Minimum frequency: 1

Maximum number of items in a grammatical relation of the common block: 25

Maximum number of items in a grammatical relation of the exclusive block: 12



明星/演員 chinese_all_trd freq = 23923/23213

[change options](#)

Common patterns

明星	21	14	7	0	-7	-14	-21	演員
----	----	----	---	---	----	-----	-----	----

a modifier	1729	4116	1.7	4.2
著名	84	738	39.0	70.4
武打	65	14	62.6	31.1
老牌	42	110	44.5	56.4
資深	6	237	11.2	53.5
知名	29	232	26.7	52.4
當紅	43	30	49.4	38.7
大	468	19	46.6	5.1
年輕	8	136	13.3	43.8
最佳	27	135	23.2	39.5
小	38	232	20.5	37.9
老	21	112	19.6	35.5
眾多	44	16	33.8	17.5
一流	9	22	18.5	25.4
名	11	198	4.6	25.3
新一代	12	13	22.9	20.6
已故	11	7	22.5	14.9

measure	688	1532	1.1	2.5
位	310	576	50.3	54.0
名	79	617	26.4	50.4
批	13	39	17.3	25.7
個	79	159	21.3	23.9
屆	13	7	12.6	5.7
場	11	5	12.5	4.9
次	9	17	8.1	9.5
possessor	459	921	1.0	2.1
知名度	6	10	17.6	21.1
國家	9	26	4.8	8.6
and/or	336	1517	0.4	1.7
導演	11	381	21.3	68.4
歌星	16	50	32.2	42.6
歌手	7	50	17.0	34.8
藝術家	6	36	15.5	30.3

n_modifier	13964	9897	1.6	1.2
大牌	162	13	57.7	21.6
偶像	297	8	57.1	11.7
喜劇	50	144	33.8	52.8
影視	248	64	50.8	32.3
好萊塢	190	70	50.5	37.0
新生代	25	133	24.0	50.3
男	41	355	16.1	46.6
演技派	31	37	39.8	44.5
歌仔戲	11	83	15.4	42.6
芭蕾舞	7	63	13.6	42.5
青年	26	564	5.4	41.7
舞蹈	12	214	7.9	41.4
電影	343	301	38.5	39.4
歌劇	11	64	13.4	35.2
男女	28	112	14.1	31.7
女	234	18	30.9	7.2

"演員" only patterns

possession 693 1.6

演技	15	33.7
演出	18	21.9
手	21	21.8
功力	6	20.8
服裝	13	19.6
積極性	9	17.6
特質	5	15.9
表現	13	15.4
肢體	5	14.9
角色	8	14.4
技巧	5	13.8
動作	6	12.3

subject_of 3924 1.6

謝幕	20	44.9
擔綱	33	41.7
親切	23	30.1
跳起	11	29.5
演戲	9	29.4
演	22	29.1
握手	19	28.2
齊聚一堂	14	27.3
載歌載舞	10	26.3
清唱	8	24.4
拍戲	6	24.3
精湛	13	24.1

n_modifier 9897 1.2

京劇	246	53.7
一級	424	50.8
相聲	120	50.4
實力派	71	49.1
雜技	129	48.5
替身	33	43.7
特技	92	43.1
特型	21	42.4
越劇	45	41.0
話劇	59	38.1
舞台劇	45	37.4
歌舞伎	21	36.9

modifies 8784 1.0

郎雄	51	50.1
李康生	29	42.4
柯俊雄	29	42.3
金士傑	22	39.9
柯受良	21	37.7
胡軍	18	36.6
姜昆	19	36.0
鄭亞雲	15	35.9
楊貴媚	20	35.3
戴立忍	19	34.2
楊呈偉	14	33.9
濮存昕	13	33.8

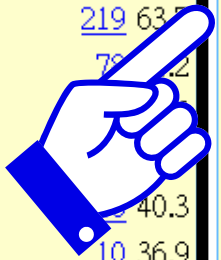
object_of 1535 0.7

獨唱	50	54.1
客串	22	41.9
演	23	34.0

"明星" only patterns

a_modifier 1729 1.7

超級	219	63.5
演藝	78	45.2
美式	10	36.9
耀眼	10	36.9
閃亮	10	36.9
過氣	10	36.9
超冷	5	30.6
三棲	6	29.8
佳	21	29.3
頭號	19	27.0
溫	12	24.9
簡	10	24.6



n_modifier 13964 1.6

籃球	433	46.3
職籃	179	45.6
足球	675	44.9
職棒	232	43.0
卡通	99	41.9
夢幻	68	40.8
網球	293	39.2
抗癌	50	34.9
偶像級	16	34.7
恆康	15	34.7
演藝圈	36	33.8
演藝界	34	33.8

modifies 13046 1.4

球員	1095	59.7
對抗賽	346	59.3
聯隊	301	54.2
排名賽	108	49.0
辛浦森	56	48.4
籃球隊	167	46.9
馬拉杜納	30	43.0
白隊	47	41.2
後衛	141	40.0
前鋒	149	38.8
馬拉多納	39	38.4
王貞治	30	37.8

possession 506 1.1

風采	26	37.4
架子	12	32.2
架勢	10	32.1
搖籃	12	30.2
名字	12	24.4
架式	5	23.8
丰采	5	21.8
光環	5	20.2
魅力	7	18.7
照片	9	18.5
故事	7	16.3
青少年	6	10.7

measure 688 1.1

類	79	51.3
隼	7	15.8
路	6	10.8
所	15	10.2
家	7	8.2

possessor 459 1.0

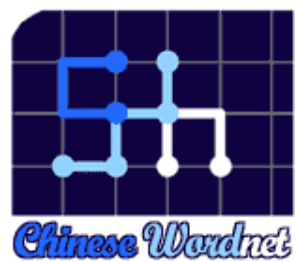
世界級	11	25.3
名氣	5	19.9
比賽	21	17.1
聯隊	6	16.9
氣	5	14.3

subject_of 2280 0.9

薈萃	53	49.3
雲集	36	39.5
三缺一	8	37.4
開店	12	34.6
評選	26	32.8

object_of 1536 0.7

啓	44	48.8
崇拜	19	34.0
棲	11	33.2
考上	16	31.8
服務	93	29.0



Automatic Discovery of Transliteration Variants: Our Recent Works

- *Huang, Chu-Ren, Petr Simon, and Shu-Kai Hsieh. 2007. Automatic Discovery of Named Entity Variants. Proceedings of the Association of Computational Linguistics Annual Meeting, Prague-Czech, June 25-28.*
- *Šimon, Petr, Chu-Ren Huang, Shu-Kai Hsieh, and Jia-Fei Hong, 2007. Transliterated Named Entity Recognition Based on Chinese Word Sketch. Proceedings of Chinese Lexical Semantics Workshop 2007, Hong Kong Polytechnic University, May 20-23*



How to judge which two different in forms are the same?

異中如何求同？

You shall know a word by the company it keeps.

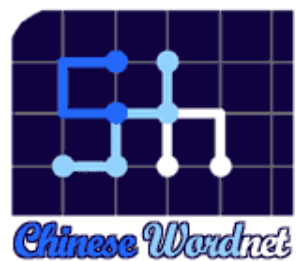
-J. R. Firth

- A named entity shall be known by the company it keeps
- All transliteration variants refer to the same name entity
- Two transliteration variants must share the same group of company
- We start with pairs of known transliteration pairs and look for mappings between their respective groups of companies (companies of 柯林頓 vs. companies of 克林頓)



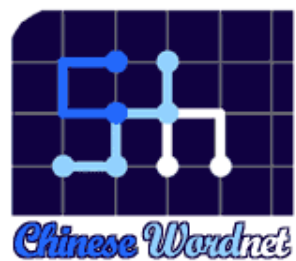
In Other Words

- Same entity is likely to occur in the same context across in comparable texts
- Different, but phonologically similar, words which occur in the same context are likely to be transliteration variants
- We use the and/or relation in WordSketch
- Relation and/or is defined as a relation of two nouns separated either by a conjunction or by an Chinese comma "、".



The Algorithm

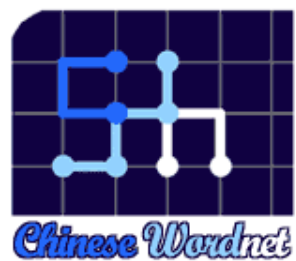
- have a set of seed pairs, each pair $S_i = \langle hwXIN ; wCNAi \rangle$. E.g. $\langle \text{克林頓}, \text{柯林頓} \rangle$
- For each seed pair, retrieve Word Sketch difference for and/or relation, thus have two word lists, $L = \langle hWXIN ; WCNAi \rangle$,
 - where Wki is an unordered list of words.
- Process each list of candidates L with the pairs extraction algorithm.



Some Possible Transliteration Variations

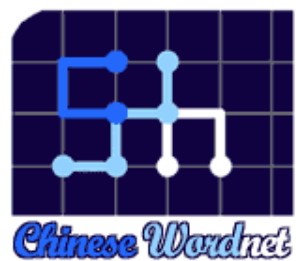
- Syllabification (or not) of ending consonant
 - Arafat, 阿拉法 vs. 阿拉法特
- Choice of gender specific characters
 - Leslie 萊絲莉 vs. 萊斯利
- Whether to spell out first name or not
 - Venus Williams 大威廉絲 / 維.威廉絲.
- Phonological interpretation,
 - Rafter 賴夫特 vs. 拉夫特
 - Connors 康諾斯 vs. 康那斯.
- Pronunciation according to original language or English:
 - Escudero 艾斯庫德 vs. 伊斯庫德

語音



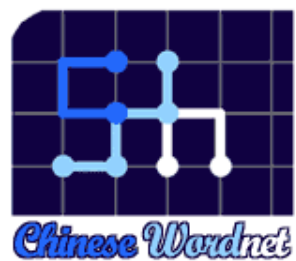
Which possible pairs are transliteration variants

- Only pairs have the same 'company' are selected
- Pairs with closest edit distance are considered potential variants
- 高爾 vs. 戈爾 – g,ao,',er vs. k,e,',er
- levenstein_distance ([g,ao,',er], [k,e,',er])
- weight initials and finals
- if score < threshold: add to seeds



Synopsis of Methodology

1. Prepare seeds from different domains
2. Retrieve Word Sketch difference - pair of lists of candidates
3. Compare phonological similarity between candidates from both lists
4. Add new pairs to seeds
5. Stop when no more new seeds are generated



The Original Seeds

XIN	CNA	English
克林頓	柯林頓	Clinton
巴赫	巴哈	Bach
喬丹	喬登	Jordan
達文西	達芬奇	Da Vinci
畢加索	畢加索	Picasso
碧咸	貝克漢	Beckham
萊溫斯基	呂茵斯	Lewinsky
阿斯平	亞斯平	Aspen
侯賽因	胡笙	Hussain
卡斯特羅	卡斯楚	Castro

語音

Word Sketch Difference: Exclusive patterns

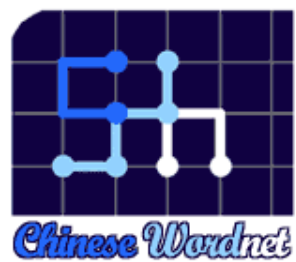
克林頓			柯林頓		
Rel	Freq1	Sal1	Rel	Freq1	Sal1
and/or	1197	1.9	and/or	3940	1.5
葉利欽	169	60.9	呂茵斯基	211	68.7
戈爾	57	51.8	高爾	311	65.9
巴拉克	48	46.0	葉爾勤	317	59.2
布什	86	43.8	布希	357	52.5
阿薩德	32	41.5	希拉蕊	86	50.1
希拉克	38	39.1	巴瑞克	62	45.8
萊溫斯基	11	34.9	宋嘉斯	26	42.9
科爾	22	33.5	阿塞德	38	40.5
侯賽因	18	31.9	裴洛	39	40.0
...



Result From First Experiment

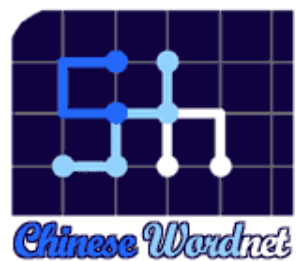
- 11 seeds
- 505 iterations
- 494 pairs extracted
- precision 90%

語音



From Transliteration Variation to Sociolinguistic Study

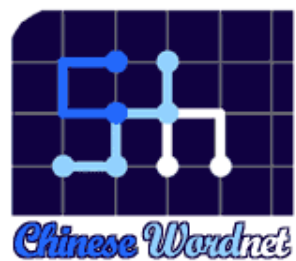
- Compare corpora from three societies
 - Mainland China, Taiwan, and Singapore
- Control and contrast seeds by different domains
 - Arts, Location, Music, Politics, and Sports



Precision of Prediction by Domains

Domain	correct	incorrect	total	precision
Art	5	0	5	100%
Music	16	0	16	100%
Location	35	228	263	13.3%
Politics	207	180	387	53.48%
Sports	36	7	43	83.72%
Total	299	450	749	40%

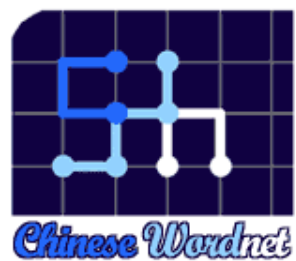
(The total number of incorrect results includes with 35 additional pairs with mismatched domains.)



Overall Results

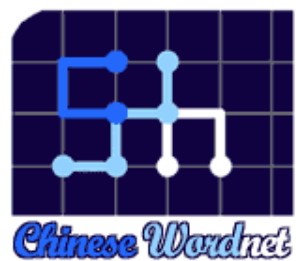
- 21 original seed pairs
- 15 effective seed pairs
 - 6 literature seed pairs yielded no results
- 449 correction predictions among 749 recalled pairs
- Art and Music, not highly productive but highly precise
- Location/Place Name: very low precision





Contrast in Three Societies

- Art/Music Domain show very little variations as they share they follow the same borrowed western tradition (Bach is the most popular among the named entities chosen.)
- Place names have the highest range of variations
- Transliteration from the domain of sports are generated even though there is no seed in that domain



Sports Domain

- Why sports: perhaps a mixture of politics and 'star' power
- Populated mostly by NBA stars (probably because of the influence of Yao Ming. Thus more news on NBA than on any other sport.)
- Samaranch is the most frequent name in this domain in PRC, but not in either Taiwan or Singapore (most likely because of national efforts to bid for Beijing Olympics)

語音

PRC	Singapore	Taiwan
意大利 Italy (2	義大利 2
克林頓 Clinton (USA)	1	柯林頓 1
葉利欽 Yel'tin (Russia)	(?)	葉爾勤/辛 3
克羅地亞 Croatia	12	克羅西亞 8
普京 Putin (Russia)	3	蒲亭 7
貝爾格萊德 Belgrade (Serbia)	27	伯爾格勒 16
塞浦路斯 Cyprus	25	塞普路斯 77
希拉克 Chirac (France)	10	席哈克 6
坦桑尼亞 Tanzania	29	坦尚尼亞 33
杜馬 Duma (France)	18	杜馬 16
佩雷斯 Perez (Israel)	17	培瑞斯 48

word	Freq.	rank	Fre/PRC	Corres. Rank/PRC	Sing. Rank
柯林頓 Clinton	120842	1	0	2	1
義大利 Italy	58245	2	0	1	2
葉爾勤 Yeltn	32338	3	0	3	46
波士尼亞 Bosnia	19171	4	0	57	15
柯索伏 Kosovo	9497	5	0	132	45
席哈克 Chirac	9176	6	1	7	10
蒲亭 Putin	9060	7	0	5	3
克羅西亞 Croatia	7948	8	1	4	12
柯爾 Khol	6221	9	669	11	9
斯多福 Stov	4659	10	0	17	36
裴利 Perry	4498	11	0	42	39
倫斯斐 Rumsfield	4305	12	0	138	46
米洛塞維奇 Milosevic	4128	13	8	16	5
施若德 Schroed	3917	14	0	13	8
葛林斯潘 Greenspan	3747	15	0	25	13

Singapore

克林頓	
意大利	
普京	
伯格	
米洛捨維奇	Milosevic (Serbia Politician)
賴斯	Rice (US Secretary of State, Female)
奧尼爾	O'Neil (US Basketball player)
施羅德	
科爾	
希拉克	
卡洛斯	(King Carlos of Spain)
克羅地亞	
格林斯潘	(Greenspan)
艾弗森	Iverson (US Basketball player)
福克斯	Fox (Mexican President)

- Note that Singapore data covers a much shorter time and hence is more sensitive to current events



Summary

- **Comparable corpora provides a way to do global comparative study of transliterated terms among different Chinese speaking societies.**
- **Domain and contextual information is generally very effective for identifying transliteration variants,**
- **However, place name is a domain where native and transliterated names cannot be differentiated by contextual collocation**

Conclusion

- *This first attempt towards research on ‘corpus-aided armchair sociolinguistics’ is only partially successful*
 - *Fillmore invented the term ‘corpus-aided armchair linguistics’ in 1992, at the beginning of the field of corpus linguistics*
- *Yet Promising. With more than 800 transliteration variants identified to attest contrastive use among different Chinese speaking communities and Gigaword Corpus to give contextual information.*