From Linguistic Insights to Theoretical Relevance: A personal retrospective on Chinese grammar

語料觀察與理論詮釋：中文語法研究的個人回顧

Chu-Ren Huang 黃居仁

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

Short English Abstract

The road from linguistic insights to theoretical relevance can be confusing. On one hand, it is easy to be lost in linguistic complexity and miss the theoretical relevance. On the other hand, it is also easy to be too fixated on a theoretical model and miss a more elegant representation of the linguistic facts. In this talk, I will present three sets of Chinese data in light of different linguistic theories. The facts studied are Primary/Secondary predicate, Subcategorized Topic, and Expletive object. The theories discussed include GPSG, LFG, and Construction Grammar. The perspective is personal since I worked on all these facts and grammatical theories. It is hoped, however, that this personal retrospective will shed some light on how to achieve synergy between linguistic insights and theoretical relevance.

中文摘要：

語言學理論的發展，自杭士基 Chomsky 於 1950 年代後期，首創衍生變形語法理論後，產生了重要的範典更新（paradigm shift)。新範典產生之後，隨之而來的是不同理論的辯證與競和。在 50 到 60 年代，理論的討論，焦點在結構主義與衍生語法間的優劣利弊。進入 70 年代以後，衍生學派作為語言學理論主流的地位已毫無置疑,並建立了語言學理論歷史上前所未有的規範理論基礎。這個新的規範理論範典，使得衍生學派中，在對人類語言的解釋不同假設時，不但可以提出不同理論架構，更可以直接對理論架構作科學性與規範性的檢視。在這個環境下，由 1980 年代到 2000 年間，語言學理論可以說是百花齊放，百家爭鳴。而且每個理論都能對人類語言的科學解釋，提供新觀點與新方法。有過相當影響力的理論，包括了 Generative Semantics, Case Grammar, Relational Grammar, GPSG, LFG 等。這 20 年間，正好是個人專心耕耘語法理論與中文語法研究的時間，對大多數的理論都有所涉獵。藉著對這些研究的回顧與整理，本文希望能釐清語料觀察與理論詮釋的關係，並從而凸顯這個過程中的一些應避免的盲點。最後，我將對語言系統性的本質，提出自己的看法。

語法理論的快速蓬勃發展，雖然對理論解釋與發展有正面的誘導；但也造成了一個危機。由於理論的更新與發展過於快速，反而不能先在大量多樣的不同語言中得到驗證。因此，理論的發展與描述，常常是取決於少數幾個當時最熱門的

語言中的局部關鍵性語料。有志之士看到了這個問題，提出呼籲，希望新理論應該針對不同語言作更全面深入的分析再提出解釋，而不是急就章的用少量單一語言語料提出新理論。在方法論上的論證，就是單一語法現象的跨語言一致性解釋，與單一語言系統性的一致解釋，到底孰先孰重？在研究的實踐上，沒有學者能夠真正把單一現象在世界上所有語言中的表現都分析完；也鮮少有學者能夠寫出涵蓋單一語言所有現象，又具嚴謹理論系統的語法。因此，兩種策略，都是逼近科學事實的研究方法，不相排斥。我個人選擇單一語言的系統性解釋，主要是認為這個方向可以做出目前比較能有具體可驗證結果的研究。

個人的理論語言學研究就是在這個背景下進行的。由 80 年代中期起，黃居仁參與了 GPSP (Generalized Phrase Structure Grammar)與 LFG (Lexical-functional Grammar) 等兩個理論的發展。這兩個理論，當時率先提出表面結構為主，詞彙驅動等與變形語法不相容的解釋觀點。對後來理論與計算語言學都發生了深遠的影響。個人率先將這兩個理論引入中文語言學研究的學者。一方面用中文來檢驗新理論的強健性(robustness) 與完整性。一方面藉由理論的新觀點，誘發中文特殊結構的分析，不但發掘新語料，也提出新解釋。在理論上，提出了架構修正的貢獻；在語料上，發掘了許多過去缺少理論解釋的中文現象（如 Expletive object, Possessive objects, Pseudo-possessive constructions, Subcategorized Topic, 等等）。這個結合新理論與新語料的研究方向，後來也延伸到 1990 年代末期興起的句構語法(Construction Grammar) 的中文研究。在這三個理論中，個人的研究都是各理論裡，最早用於中文語法分析的文獻。在這些論文中，也都能採理論與語言多樣性兼顧的立場，做出兼具理論改進與新語料解釋的成果。換句話說，這些研究不但是個別理論用於中文的尖端研究，更是新理論與新語料互動，激發新觀點的例子。

本文終將針對三組中文特有的語言現象：Primary/Secondary predicate, Subcategorized Topic, 及 Expletive object；探討語料觀察, 系統性的描述分析，與理論詮釋間的平衡點。期望能見微知著，導出語言為知識系統的新觀點，而給予語言學理論更宏觀而完備的解釋能力。

References

I. Grammatical Theories

Bresnan, Joan. 1982. Ed. *The Mental Representation of Grammatical Relations*. Cambridge: MIT Press.

Gazdar, Gerald, Ewan Klein, Geoffrey Pullum, and Ivan Sag. 1985. *Generalised Phrase Structure Grammar*. Oxford: Blackwell.

Goldberg, Adele. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.

II. Chinese Grammar and Linguistic Theory: a personal bibliography

Huang, Chu-Ren. 1985. On Pseudo-Possessive NPs in Mandarin Chinese. *Cornell Working Papers in Linguistics.* 7.129-148.

Huang, Chu-Ren and Louis Mangione. 1985. A Reanalysis of de: Adjuncts and Subordinate Clauses. *Proceedings of WCCFL IV.* 80-91. Stanford: Stanford Linguistics Association.

Huang, Chu-Ren. 1986. Coordination Schemas and Chinese NP Coordination in GPSG. *Cahiers de Linguistique Asie Orientale.* XV.1.107-127.

Huang, Chu-Ren. 1988. Head-wrapping and Possessive Objects - Syntactic vs. Lexico-semantic Treatment of Discontinuity. *Tsing Hua Journal of Chinese Studies.* New Series 18.1.161-177.

Huang, Chu-Ren. 1988. Towards a Morphosyntactic Account of Taiwanese Question Particle *kam*. *Bulletin of the Institute of History and Philology*. 59.425- 447.

黃居仁. 1988. 再析國語「領屬主語」結構─概化詞組結構語法與詞彙功能語法之比較研究。（Possessive Subjects Revisited -- A Comparative Study of GPSG and LFG.） 漢學研究 *Sinology Studies.* 6.2.143-168.

黃居仁 1989. 聯併(Unification): 語法理論與剖析。Proceedings of ROCLING I. 29-54.

Shiu, Yu-Ling and Chu-Ren Huang. 1989. A Unification-based Approach to Mandarin Questions. *Journal of Information Science and Engineering*. 5.4.395-420.

Huang, Chu-Ren. 1989. Subcategorized TOPIC in Mandarin Chinese. The Chinese Language Teachers Association Annual Conference. Boston. Nov.17-19th.

Huang, Chu-Ren. 1990. Review. [J. Tai and F. Hsueh (eds.), 1989, Functionalism and Chinese Grammar.] *Journal of Chinese Linguistics.* 18.2.318-334.

Huang, Chu-Ren. 1991. Adjectival Reduplication in Southern Min -- A Study of Morpholexical Rules with Syntactic Effects. *Chinese Languages and Linguistics*: Volume. I. (Chinese Dialects). Pp. 407-422. Taipei: Academia Sinica.

Huang, Chu-Ren. 1992. Certainty in Functional Uncertainty. *Journal of Chinese Linguistics*. 20.2.247-287.

Huang, Chu-Ren, and Fu-wen Lin. 1992. Composite Event Structures and Complex Predicates: A Template-Based Approach to Argument Selection. In Laurel S. Stvan et al. Eds. FLSM III: Papers from the Third Annual Meeting of the Formal Linguistics Society of Mid-America. Pp. 90-101. Bloomington: IULC Publications.

Huang, Chu-Ren. 1992. Certainty in Functional Uncertainty. *Journal of Chinese Linguistics*. 20.2.247-287.

Huang, Chu-Ren. 1993. Reverse Long-distance Dependency and Functional Uncertainty: The Interpretation of Mandarin Questions. In Chungmin Lee and Boem-mo Kang (Eds.) *Language, Information, and Computing*. Seoul, Thaehaksa. pp. 111-120.

Chen, Chao-ran, Chu-Ren Huang, and Kathleen Ahrens. 1995. Construction as a Theoretical Entity: An Argument Based on Mandarin Existential Sentences. *Proceedings of the 10th Pacific Asia Conference on Language, Information, and Computing.* City University of Hong Kong. December 27-28th.

Huang, Chu-Ren 1997. Morphological Transparency and Autonomous Morphology: A Comparative Study of Tough Constructions and Nominalization. *Chinese Language and Linguistics III: Morphology and Lexicon.* Pp.369-399. Taipei: Institute of History and Philology, Academia Sinica.

Huang, Chu-Ren, Zhao-ming Gao, Kathleen Ahrens, and Wen-Jen Wei. 1998. Argument Structure and Empty Expletive: A study of Event Evaluative Predicates in Chinese. In Bejamin T'sou (Ed.) *Studia Linguistica Serica.* Pp. 401-413. Hong Kong: City University of Hong Kong.

Huang, Chu-Ren, Li-Ping Chang, Kathleen Ahrens, and Chao-Ran Chen. 1999. 詞彙語意和句式語意的互動關係。（The Interaction of Lexical Semantics and Constructional Meanings）. In Y. Yin et al. Eds. *Chinese Language and Linguistics V: Interactions in Language.* Pp.413-438. Taipei: Institute of Linguistics, Academia Sinica.

Chief, Lian-Cheng, Chu-Ren Huang , , Keh-Jiann Chen, Mei-Chih Tsai, and Li-Li Chang. 2000. What Can Near Synonyms Tell Us? InYung-O Biq. (Ed.) Special Issue on Chinese Verbal Semantics. *Computational Linguistics and Chinese Language Processing.* 5.1.47-60.

From Linguistic Insights to Theoretical Relevance
語料觀察與理論詮釋

Chu-Ren Huang
Academia Sinica
http://cwn.ling.sinica.edu.tw/huang/huang.htm

Or: How to avoid getting deleted?

Old linguists don't die, they just get obligatorily deleted.

-Title of Eva Hajicova's talk given at her speech when accepting the Association of Computational Linguistics Life Achievement Award in July 2006

-From an old Indiana University Linguistics Club T-Shirt

How to avoid getting deleted I

- **I have data, lots of data!**

And some linguistic generalizations too.

How Much Data? Comparing Chinese Corpora

Corpus Name	Online Year	Data : million words/characters	Duration/ Content
Sinica 4.0 (Taiwan)	1996	5.2 M words 7.9 M characters	1990-1996 Fully Tagged
Sinica 5.0 (Taiwan)	2006	10 M words	1990-2004 Fully Tagged
Sinorama (Taiwan)	2003	3.2 M English words 5.3 M Chinese characters	1976 – 2000 (1999-2000) Aligned
CCL (Peking)	2003	85 M simplified characters	1919 -2003 Partially tagged (1 million)

How to find generalizations from Corpus I

- Size Matters
  - Observable facts are directly proportional to corpus size
  - An average collocation requires corpus size of 1,000,000,000 (10,000x10,000x10)

    Hence the Gigaword Corpus

How Much Data Can One have? 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

Size of Chinese Gigaword Corpus

	Resource	Characters/ million	Words/ million	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: **Million**

Size of Lexicon Extracted from Chinese Gigaword Corpus

	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

How to find generalizations from Corpus II

- Numbers (and tendencies) don't lie
  - A large enough sample allows us to observe the tendencies (distributional patterns) of language use
- Linguistic Knowledge can be used to generate more linguistic knowledge
  - In order to extract grammatical relations automatically and efficiently, we applied detail lexical grammatical knowledge, especially argument structures and collocating patterns of verbs

Finding a Word's Company: Corpus KeyWord In Context (KWIC) and the color pen approach

arity, which will be used to take a party of under-privileged children to
2 from outside. You are invited to a party and after a couple of drinks you
3 tion, we believe politicians of all parties will listen to our views. &eq
4 ould be reaching agreement with all parties concerned as to which event
5 lack people. I have certainly been party to one or two discussions amongs
6 . These should be discussed by both parties before entering into the relat
7 presents They had hosted a cocktail party at Kensington palace, for examp
8 akes. By midnight the end-of-course party is in full swing, but most cad
9 e should be a right for the injured party to terminate the contract. A ma
10 by the Safran Peoples ' Liberation Party. This presents the powerful nei
11 s. Ahead I could see the rest of my party plodding towards the final slope
12 cial ethic. The two main political parties - the Tories and the Liberals
13 ritish successes in Perth The small party of British players competing in
14 to help control. One member of the party went to summon the rescue team a
15 rket society fashion magazine. The party was held at his flat which was a
16 security and secrecy than any Tory Party Conference : it seems that bootl

1 political association
2 social event
3 group of people
4 person in an agreement/dispute
5 to **be party to** something...

The coloured pens method — from Kilgarriff et al. 2005

A Word's Company Automatically Detected: WordSketch w BNC Data

Home | Concordance | Word Sketch | Thesaurus | Sketch-Diff

speak BNC freq = 24858

change options

object	3013	1.9	subject	4183	5.3	modifier	5463	6.8	and/or	558	0.3
english	412	57.8	non-english	7	23.67	strictly	245	60.0	write	114	38.76
language	382	46.73	english	48	23.32	broadly	180	55.67	listen	29	29.8
french	132	45.5	voice	72	21.06	generally	329	52.72	move	29	21.62
word	370	41.78	Jesus	23	18.26	again	263	42.03	act	18	20.98
truth	115	38.1	God	37	17.49	softly	94	41.88	sing	12	19.56
spanish	23	36.15	führer	7	16.11	quietly	108	41.01	speak	16	18.34
italian	37	31.46	man	96	14.97	roughly	61	38.55	understand	15	16.88
gaelic	15	31.05	Silas	7	14.88	figuratively	17	36.05	come	26	14.66
german	53	30.98				ill	28	35.77	read	11	13.8

Sketch Engine and Chinese WordSketch

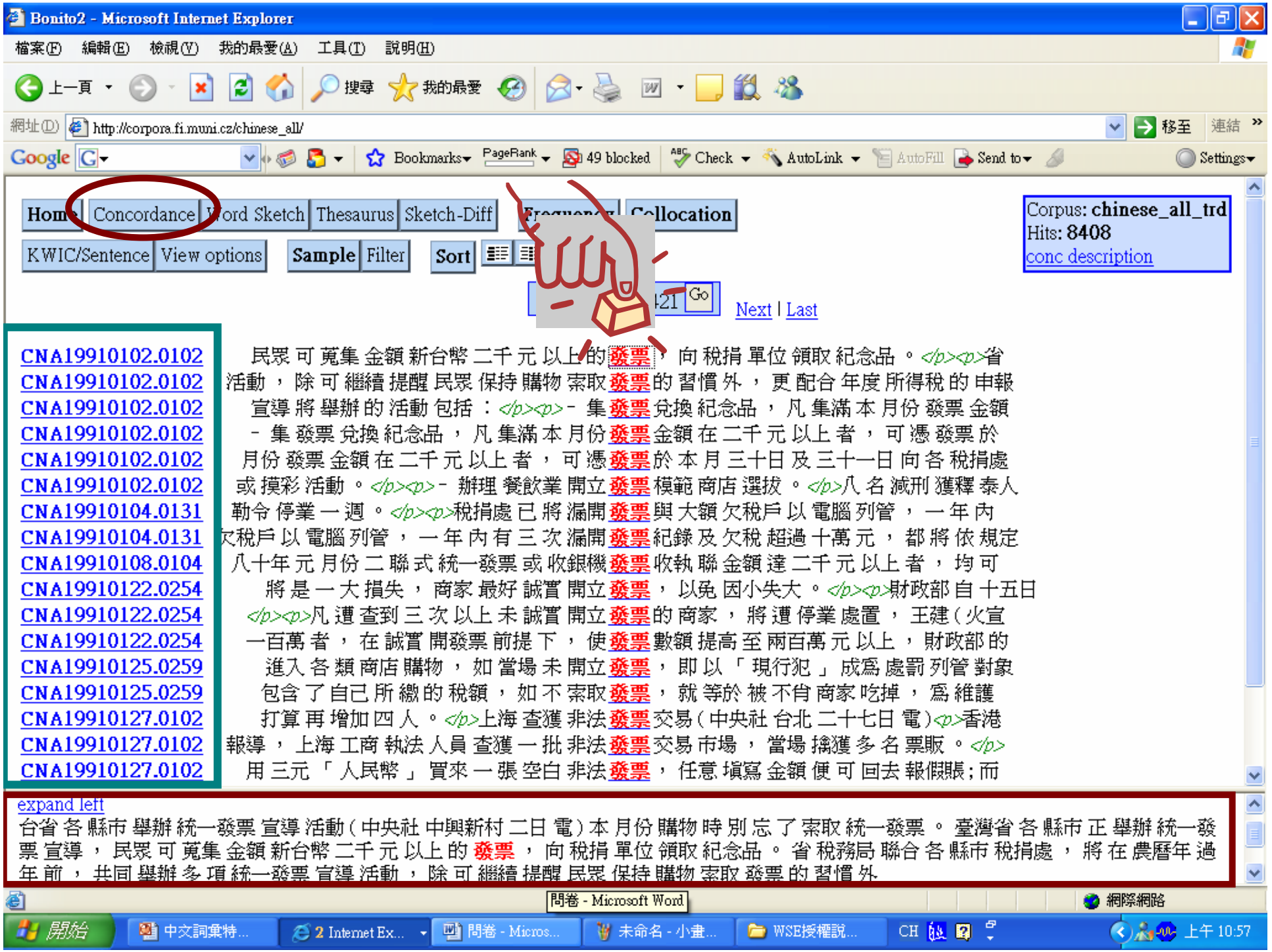
- Sketch Engine http://www.sketchengine.co.uk

Developed by team led by Adam Kilgarriff

- **A new corpus viewing tool**
- **Discovering grammatical information from a gigantic corpus**
- Chinese Wordsketch by Academia Sinica

http://www.ling.sinica.edu.tw/wordsketch (free license for use in Taiwan only)

  - **Academia Sinica, Taiwan** (Huang, Smith, Ma, Simon 黃居仁，史尚明，馬偉雲，石穆)


Bonito2 - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛
網址(D) http://corpora.fi.muni.cz/chinese_all/
移至 連結
Google
Bookmarks PageRank 49 blocked Check AutoLink AutoFill Send to Settings
Home Concordance Word Sketch Thesaurus Sketch-Diff Frequency Collocation
KWIC/Sentence View options Sample Filter Sort
Corpus: chinese_all_trd
Hits: 8408
conc description
Go
Next | Last
CNA19910102.0102 民眾可蒐集金額新台幣二千元以上的發票，向稅捐單位領取紀念品。 省
CNA19910102.0102 活動，除可繼續提醒民眾保持購物索取發票的習慣外，更配合年度所得稅的申報
CNA19910102.0102 宣導將舉辦的活動包括： - 集發票兌換紀念品，凡集滿本月份發票金額
CNA19910102.0102 - 集發票兌換紀念品，凡集滿本月份發票金額在二千元以上者，可憑發票於
CNA19910102.0102 月份發票金額在二千元以上者，可憑發票於本月三十日及三十一日向各稅捐處
CNA19910102.0102 或摸彩活動。 - 辦理餐飲業開立發票模範商店選拔。 八名減刑獲釋泰人
CNA19910104.0131 勒令停業一週。 稅捐處已將漏開發票與大額欠稅戶以電腦列管，一年內
CNA19910104.0131 欠稅戶以電腦列管，一年內有三次漏開發票紀錄及欠稅超過十萬元，都將依規定
CNA19910108.0104 八十年元月份二聯式統一發票或收銀機發票收執聯金額達二千元以上者，均可
CNA19910122.0254 將是一大損失，商家最好誠實開立發票，以免因小失大。 財政部自十五日
CNA19910122.0254 凡遭查到三次以上未誠實開立發票的商家，將遭停業處置，王建(火宣
CNA19910122.0254 一百萬者，在誠實開發票前提下，使發票數額提高至兩百萬元以上，財政部的
CNA19910125.0259 進入各類商店購物，如當場未開立發票，即以「現行犯」成爲處罰列管對象
CNA19910125.0259 包含了自己所繳的稅額，如不索取發票，就等於被不肖商家吃掉，爲維護
CNA19910127.0102 打算再增加四人。 上海查獲非法發票交易(中央社 台北 二十七日 電) 香港
CNA19910127.0102 報導，上海工商執法人員查獲一批非法發票交易市場，當場擒獲多名票販。
CNA19910127.0102 用三元「人民幣」買來一張空白非法發票，任意填寫金額便可回去報假賬；而
expand left
台省各縣市舉辦統一發票宣導活動(中央社 中興新村 二日 電)本月份購物時別忘了索取統一發票。臺灣省各縣市正舉辦統一發票宣導，民眾可蒐集金額新台幣二千元以上的發票，向稅捐單位領取紀念品。省稅務局聯合各縣市稅捐處，將在農曆年過年前，共同舉辦多項統一發票宣導活動，除可繼續提醒民眾保持購物索取發票的習慣外
問卷 - Microsoft Word
網際網路
開始 中文詞彙特... 2 Internet Ex... 問卷 - Micros... 未命名 - 小畫... WSE授權說... CH 上午 10:57

WordSketch's Approach: From Lexical Types to Relations Types

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

Linguistic Knowledge in Use

Example tuple (<eat, obj, rice>)

- Challenge: Long-distance relations
- 全穀麵包，吃了很健康。

  *quan.gu mian.bao, **chi** le hen jian.kang*
- 有人嘗試要將這荷花分類，卻越分越累。

  *you ren chang.shi yao **jiang** zhe he.hua **fen.lei**, que yue fen yue lei*
- 他 只 吃了 一 口 飯 …

  Ta *zhi **chi** let yi kou **fan***

Applying Linguistic Knowledge in Processing

Knowledge Source

- Information-based Case Grammar (ICG, Chen and Huang 1992)
- Encoded on over 40,000 verbs in Sinica Lexicon
  - ICG Basic Patterns for Stative Pseudo-transitive Verb (VI)

    EXPERIENCER<GOAL[PP[對]]<VI
    EXPERIENCER<VI<<GOAL[PP[於]]
    THEME<GOAL[PP{對、以}]<VI
    THEME<VI<<GOAL[PP[於]]
    THEME<VI<<SOURCE[PP{自、於}]
    THEME< SOURCE[PP{歸、爲}]<VI

Linguistic Knowledge in Action

- 村莊(object) 明天將 被 夷爲平地(VB11)

  cunzhuang mingtian jiang bei yiweipingdi

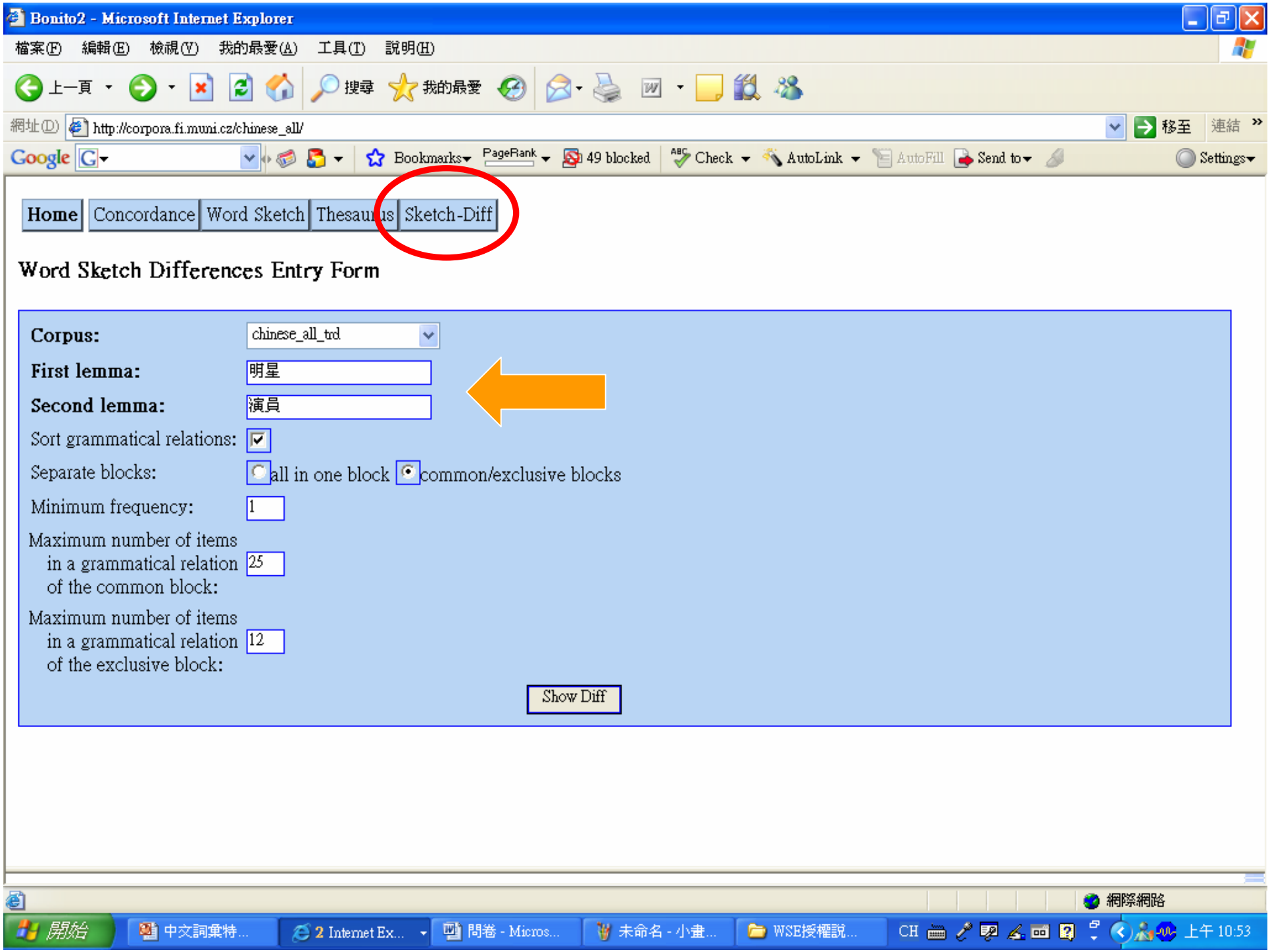
  - begin time1 location time1 adv? passive_prep adv_string 1:"V[BCJ].*" [tag!="DE"]
- 大量 的 遊客 破壞(VC2) 公園 景觀(object)

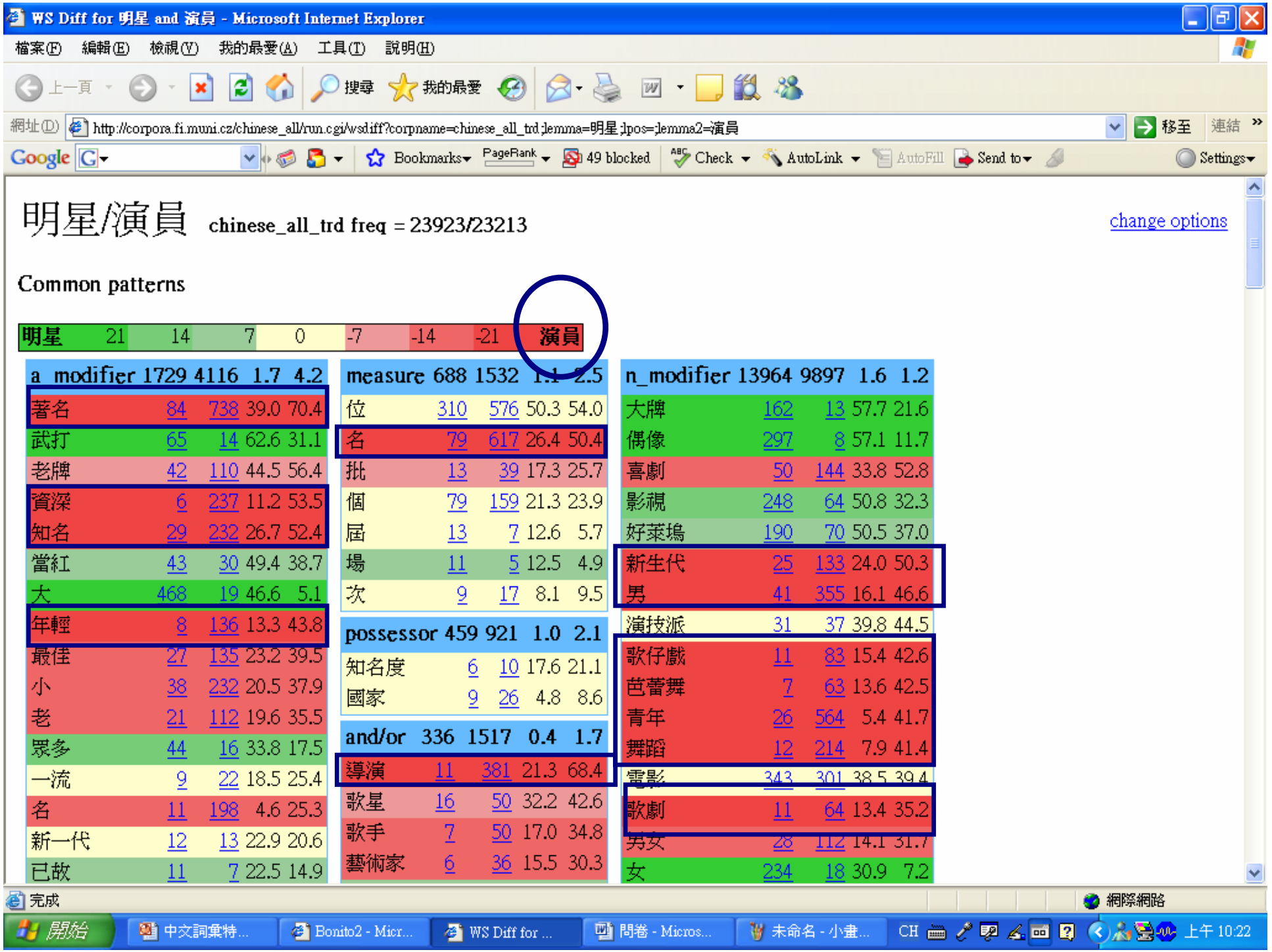
  daliang de youke pohuai gongyuan jingguan

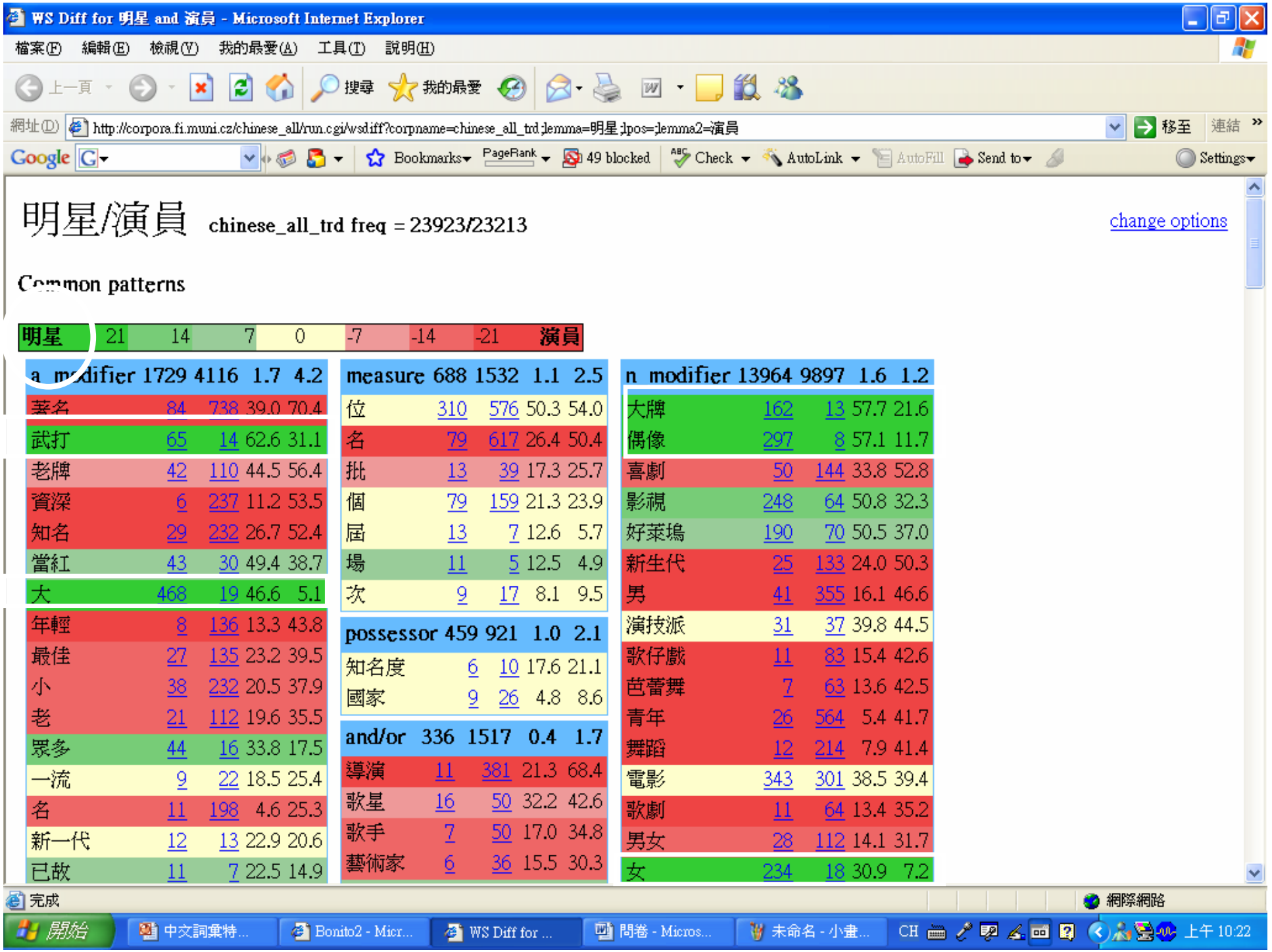
  - 1:"VC.*" (particle|prep)? NP not_noun
  - (NP is defined as "…noun_modifier{0,2} 2:noun…".

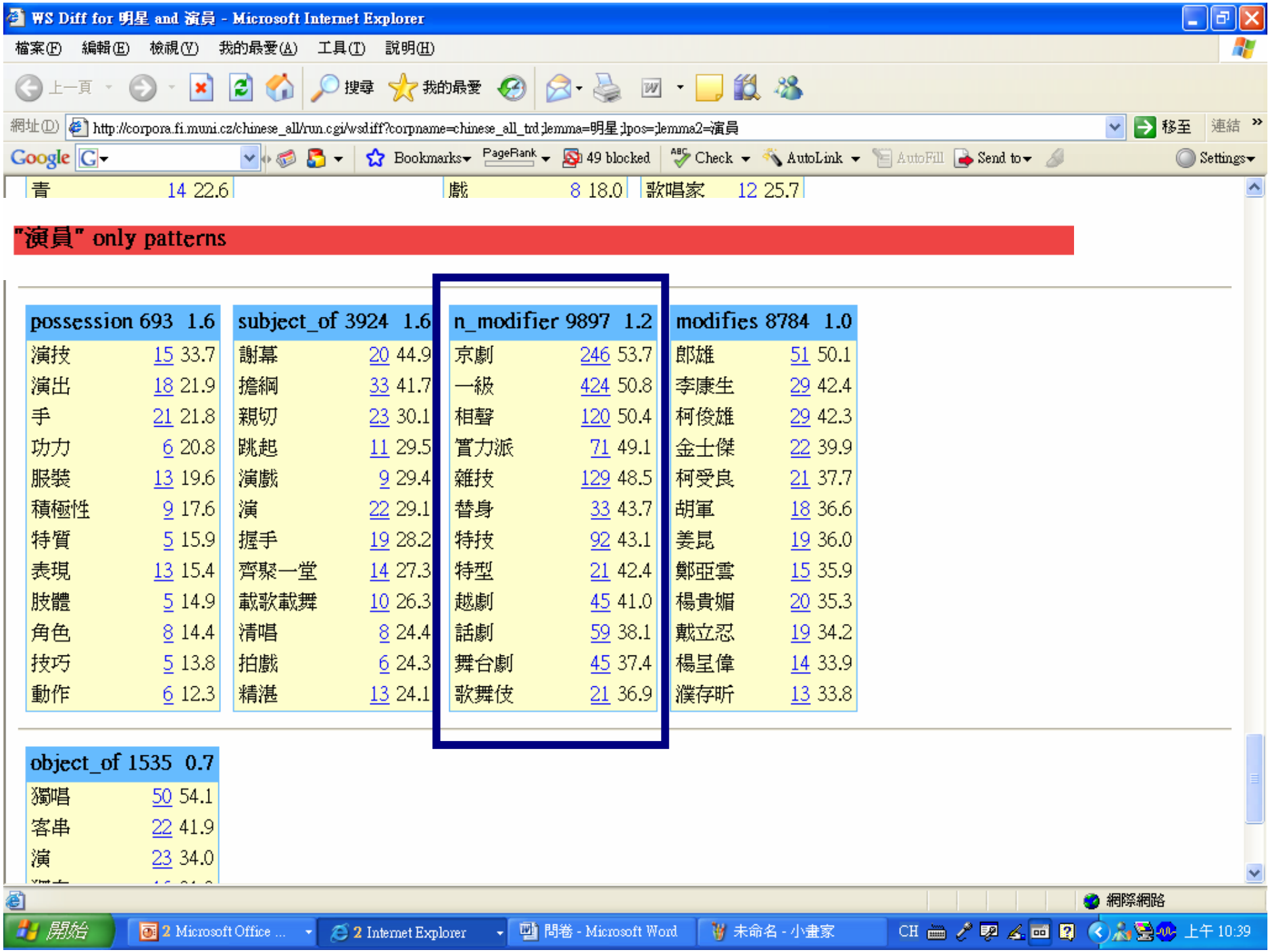
WordSketch Example:「發票invoice」

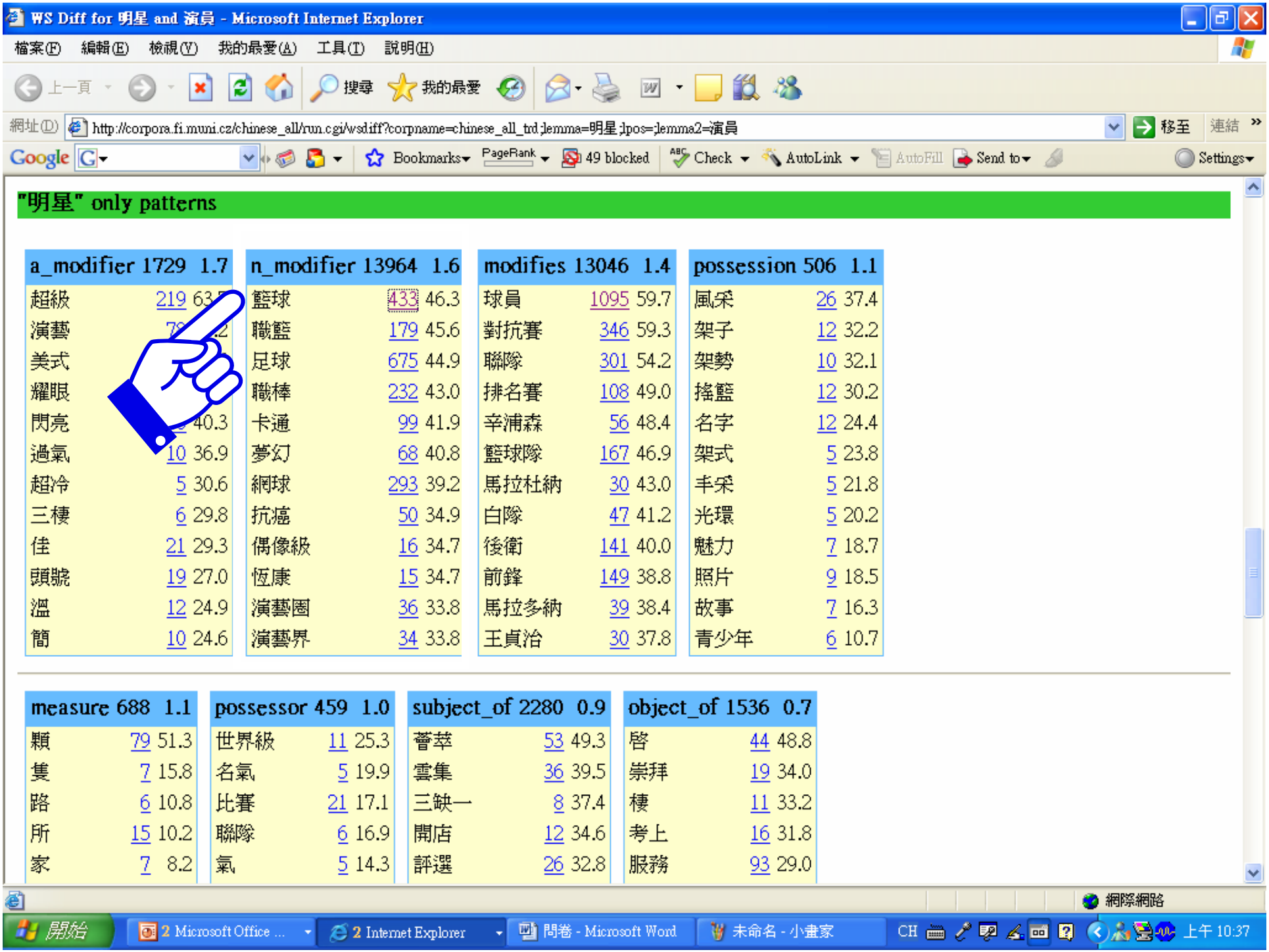
object_of 928 5.2		subject_of 692 0.3		a_modifier 276 0.5		n_modifier 562 -12.3		modifies 607 -12.8	
開立	253 73.78	收執	15 43.41	不實	85 57.41	增值稅	103 47.76	逃漏稅	16 30.08
虛開	50 51.4	給	69 32.61	假	48 46.83	銷貨	17 34.96	存根聯	7 29.53
偽造	70 40.88	對獎	7 29.59	空白	24 41.9	收銀機	16 34.19	金額	55 27.38
虛購	12 40.47	中獎	12 25.43	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


Bonito2 - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛
網址(D) http://corpora.fi.muni.cz/chinese_all/
移至 連結
Google
Bookmarks PageRank 49 blocked Check AutoLink AutoFill Send to Settings
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
Show Diff
網際網路
開始
中文詞彙特...
2 Internet Ex...
問卷 - Micros...
未命名 - 小畫...
WSE授權說...
CH
上午 10:53


WS Diff for 明星 and 演員 - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛
網址(D) http://corpora.fi.muni.cz/chinese_all/run.cgi/wsdiff?corpname=chinese_all_trd;lemma=明星;lpos=;lemma2=演員
移至 連結
Google Bookmarks PageRank 49 blocked Check AutoLink AutoFill Send to Settings
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
完成
網際網路
開始 中文詞彙特... Bonito2 - Micr... WS Diff for ... 問卷 - Micros... 未命名 - 小畫... CH 上午 10:22


WS Diff for 明星 and 演員 - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛
網址(D) http://corpora.fi.muni.cz/chinese_all/run.cgi/wsdiff?corpname=chinese_all_trd;lemma=明星;lpos=;lemma2=演員
移至 連結
Google Bookmarks PageRank 49 blocked Check AutoLink AutoFill Send to Settings
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
完成
網際網路
開始 中文詞彙特... Bonito2 - Micr... WS Diff for ... 問卷 - Micros... 未命名 - 小畫... CH 上午 10:22


WS Diff for 明星 and 演員 - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛
網址(D) http://corpora.fi.muni.cz/chinese_all/run.cgi/wsdiff?corpname=chinese_all_trd;lemma=明星;lpos=;lemma2=演員
移至 連結
Google
Bookmarks PageRank 49 blocked Check AutoLink AutoFill Send to Settings
青 14 22.6
戲 8 18.0
歌唱家 12 25.7
"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
網際網路
開始
2 Microsoft Office ...
2 Internet Explorer
問卷 - Microsoft Word
未命名 - 小畫家
CH
上午 10:39


WS Diff for 明星 and 演員 - Microsoft Internet Explorer
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
上一頁 搜尋 我的最愛
網址(D) http://corpora.fi.muni.cz/chinese_all/run.cgi/wsdiff?corpname=chinese_all_trd;lemma=明星;lpos=;lemma2=演員
移至 連結
Google
Bookmarks PageRank 49 blocked Check AutoLink AutoFill Send to Settings
"明星" only patterns
a_modifier 1729 1.7
超級 219 63.7
演藝 79 [illegible]
美式
耀眼
閃亮 40.3
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
顆 79 51.3
隻 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
網際網路
開始 2 Microsoft Office ... 2 Internet Explorer 問卷 - Microsoft Word 未命名 - 小畫家 CH 上午 10:37

Mining Cross-Strait Lexical Difference

- **Strategy: Using a pair of know contrasting words as seeds and lookup SketchDifference**
  - **Clinton 克林頓 ke4 Vs. 柯林頓 ke1**
- **What is found: Other unique translation for either PRC or Taiwan**
- **克林頓 (PRC) only and/or patterns (vs柯林頓 only)**

葉利欽88 54.6 Yeltin 葉爾勤 (3)
布什65 49.7 Bush 布希 (4)
萊溫斯基10 41.3 Lewinsky 呂茵斯基 /呂女 (1)
戈爾 20 39.4 Gore 高爾 (2)

The question is,

- What are we (linguists) going to do with so much data
- How to analyze and map it promptly

How to avoid getting deleted II

- **I have theory, fancy theory!**

And some linguistic generalizations too.

Grammatical Theories: A Selective Survey

- Early Influences
- Structuralism
- (Mainstream) Chomskyan Generative Grammars
- Other Generative Grammars
- ‘Functional’ Grammars

Grammatical Theories: Early Influences

- Rene Descarts's rationalism (what Chomsky calls 'Cartesian Linguistics'
  - Generalizations about the world can only be proved by rational reasoning (the famous *Cogito ergo sum* 'I think, therefore I am.'
- The Young Grammarian of Leipzig in the 19th century (The Neogrammarians, German *Junggrammatiker*)
  - Regularity of sound changes
  - No exception tot grammatical rules

Grammatical Theories: Structuralism

- De Saussure: The Sign/Signified Dichotomy
- American Structuralism: Jespersen, Bloomfield
  - Parallel development of Relativism
- European Structuralism
  - The Prague Circle
  - Dependency, Theme/Rheme

Grammatical Theories: (Mainstream) Chomskyan Generative Grammars

- Early Transformational Grammar
  - Derivational rules: D- and S-Structure
  - Island Conditions
- Generative Semantics
  - Kill = cause to die [with structural description]
- Government and Binding: Parameter setting
- The Minimalist Program
  - Principle and Parameter

Grammatical Theories: Other Generative Grammars

- Relational Grammar: David Perlmutter
  - 1, 2, 3 and Chomeur
- Case Grammar: Charles Fillmore
  - Agent, Patient, etc.
  - FrameNet, and Construction Grammar
- GPSG: Generalized Phrase Structure Grammar: Gazdar, Klein, Sag, Pullem

Grammatical Theories: Other Generative Grammars II

- Head-driven Phrase Structure Grammar (HPSG): Pollard and Sag
  - Derived from GPSG
- Lexical-functional Grammar (LFG): Bresnan, Kaplan
- Tree-Adjoin Grammar: Joshi
- Categorical Grammar: Montague Semantics
  - Categorical Unification Grammar

Grammatical Theories: Functional Grammars

- ?Construction Grammar: Goldberg, Fillmore, Kay
- ?Role and Reference Grammar: Van Valin
- Cognitive Grammar: Langacker

Back to the Basics: What can a grammatical theory explain

Aristotle' four causes of reason and Pustejovsky's generative lexicon

- The qualia structure of linguistic theory
  - Formal
  - Constitutive
  - Agentive
  - Telic

Linguistics: Systematic knowledge of language

- What is in the system?
  - Formal: from Sign to Structure, Structuralism
  - Constitutive: from IA to IP, rule and transformation based theories PS Grammars, TGs, Case Grammar, GPSG, LFG etc.
  - Agentive: How grammar come to being, UG approaches GB/PP, Evolution of language
  - Telic: Function and Use based Theories

Why is My Theory Deleted?

- None of the existing theory attempt to account for all aspects of language use
- Most are designed to explain only one aspect or two
- Linguists often uses arguments from one aspect to ‘delete’ other theories which are NOT designed to account for that aspect

How can we save the theories from deletion?

- There must be an underlying and unifying premise motivating all four aspects
- Which will allow synergy and integration of theories with different perspectives
- I propose that the premise is Language and a Knowledge System
  - Which is shared by all aspects

Towards Language as Knowledge System

- A complete theory to account for language as an integral sum of the four aspects.
- Atoms of knowledge：lexicalized concepts
- ‘frames’ of knowledge：lexical semantic relations
- Instantiation of knowledge：corpus

Examples of our approach to a language's knowledge system

- **Two Verbs of Ingestion: chi1吃 vs. he1喝** (Huang and Hong 2006, Hong et al. 2007)
- **The Shakespearean Garden approach to Tang Poem** (Huang et al. 2004)
- Other On-going Work
  - **Conventionalized Ontology of Chinese Writing System**
  - **From Basic Lexicon to Linguistic Ontology**

What can direct human experience tell us about conceptualization?

- Direct human experience is more concrete and involves less abstraction
  - Hence may allow more idiosyncrasies
- The introduction of a larger corpus and more powerful tool also allows us to obtain a finer grained picture of the collocational patterns
  - Beyond the liquid/solid food contrast

Data Collection and Processing

- Extract all instances of chi1 and he1 from the tagged Chinese Gigaword Corpus (*53,654 chi1, 19,561 he1*)
- Use Chinese Word Sketch to extract and rank saliency of all collocating grammatical patterns
- Use SketchDiff to compare the gram. patterns of these two verbs

Initial Observations:吃 vs. 喝

- *Chi1* takes solid objects
  - e.g. fruit, food, dinner
- *He1* takes liquid objects
  - e.g. tea, coffee, wine
- This should offer a classical case of selectional restriction
- Corpus data: many significant counter-examples
  - *chi1 nai3 shui3* (to eat milk)
  - *he1 xi1 fan4* (to drink porridge)

Chi1 'eat' only patterns

"吃" only patterns

SentObject_of 3859 7.2		
愁	103	57.9
講究	27	32.5
嚐試	13	26.5
忌	8	25.9
寧可	16	25.3
捨得	11	24.8
擔心	46	24.1
拒絕	45	22.9
寧願	11	21.2
討厭	7	21.0
忘	13	20.8
記得	12	20.3

Modifier 13757 4.5		
倒	128	41.6
津津有味	19	36.8
怎麼	78	35.0
硬	41	32.5
有得	20	31.8
不用	47	31.0
按時	35	29.4
常年	28	27.4
該	47	25.2
年年	24	24.5
一律	31	23.3
著實	13	23.3

Subject 11519 4.3		
飯	718	78.7
啞巴	30	42.0
最愛	72	39.4
柿子	27	33.5
糖	45	31.5
金飯碗	14	31.5
全家	44	31.1
魚	73	30.0
東西	78	29.4
啞吧	11	29.4
們同	10	29.2
金碗	9	27.7

Object 33038 3.7		
敗仗	326	72.3
晚飯	310	71.6
飯	802	71.0
定心丸	211	68.0
午飯	241	67.5
大鍋飯	245	66.6
閉門羹	173	66.5
年夜飯	270	65.8
狗肉	190	61.5
肉	488	60.1
虧	329	59.2
頓飯	84	59.2

PP_在 165 1.5		
工地	17	32.3

Modifies 1527 0.1		
東西	109	47.6

He1 'drink' only patterns

"喝" only patterns

SentObject_of 1065		4.9
涉	8	17.7
盛行	5	17.6
疑	5	15.0

Object 16684		4.6
花酒	666	90.8
咖啡	969	74.3
啤酒	421	58.5
下午茶	101	57.1
白開水	67	54.3
口水	114	52.5
開水	108	51.2
杯	237	50.3
飲料	283	49.6
紅酒	83	49.6
花酒案	33	49.3
綠茶	88	48.7

Modifier 4501		3.7
厲聲	13	35.5
猛	17	21.5
成天	5	19.7
獨自	7	16.8
有沒有	7	16.7
一口	5	16.2
盡情	6	15.9
其實	5	13.0
太	7	12.8
難免	5	12.7
總共	5	9.0
並	12	3.6

Subject 3235		3.0
開水	19	32.3
口水	16	30.0
胡吃海	5	28.5
馬永成	7	19.9
蕭敦仁	5	19.3
友人	17	18.9
校長	30	18.1
自來水	17	17.7
粥	5	17.6
督學	6	15.8
牛奶	7	15.8
檢察官	20	14.8

Modifies 306		0.1
水	44	34.5
飲料	15	28.6

Significant Contrast Among Shared Grammatical Patterns

吃/喝 chinese_all_trd_test freq = 53654/19561

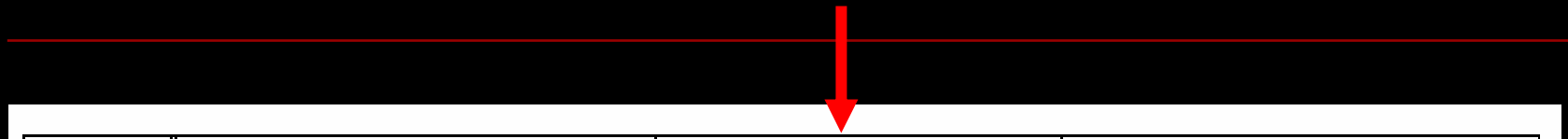
Common patterns

吃	21	14	7	0	-7	-14	-21	喝

Object	33038	16684	3.7	4.6
酒	13	5198	7.5	106.9
茶	7	825	7.0	75.5
藥	1558	8	73.0	7.3
牛奶	24	386	19.4	65.6
春酒	5	119	13.6	63.0
東西	639	16	53.3	11.1
食物	610	6	52.9	5.0
喜酒	6	43	17.7	48.8
奶	160	106	46.5	44.6
頓	167	5	43.9	8.0
稀飯	47	23	41.5	33.9
水	16	320	2.4	35.5
習慣	181	165	31.5	35.3
碗	75	19	35.2	21.4
奶水	18	20	29.8	34.3
母奶	24	20	32.3	32.7

Modifier	13757	4501	4.5	3.7
少	440	95	65.6	46.5
多	1289	304	61.7	46.6
同	384	5	52.2	7.4
不	1885	909	45.7	45.1
一起	317	159	44.1	42.0
大口	5	24	17.6	44.1
常	198	94	43.3	39.7
天天	76	23	42.8	30.9
沒	307	79	42.8	31.2
邊	145	27	41.7	25.1
連	179	20	41.1	19.4
只	363	128	37.0	31.1
不要	214	110	36.8	35.3
給他	54	10	34.9	18.9
不能	258	71	34.3	25.3
著	177	36	34.0	21.2

CWS analysis



	More usage for「chi1」	Common usage	More usage for「he1」
object	yao4 (medicine)、dong1 xi1 (foodstuff)、shi2 wu4 (foodstuff)…	xi1 fan4 (porridge)、xi3 jiu3 (wedding banquet)、nai3 shui3 (milk)、leng3 yin3 (cooling drink)…	jiu3 (wine)、cha2 (tea)、ku3 shui3 (complaints)…

Table 1: The common patterns for *chi1'eat'* and *he1'drink'*

Why Classical Selectional Restriction Theory Fails

- Selectional restriction: verb selects objects with desirable features
  - The features: [+/- food], [+/- solid], [+/- liquid]
- The collocational patterns extracted from Chinese Word Sketch suggests otherwise
- Neutralized selection: Metaphorical and metonymic extensions point to the inadequacy of a feature checking account

Xi1-Fan4 (Porridge)

- Porridge: A kind of food
- The patient: Both [+solid] and [+liquid] substances
- The role type is ambivalent
- It contains two important ingredients: Rice (solid) in soup (liquid)

Xi3-Jiu3 (Wedding Banquet)

- Wedding Banquet: A metonymical extension
- Metaphorical: Eventive and coerced
  - From entity-type patient to event-type patient
  - [+solid] and [+liquid] attributes inherited
- Include sub-events of eating solid food and drinking wine

Nai3-Shui3 (Milk)

- Interesting Fact: The neutralization of chi1/he1 milk depends on the subject. It is possible only when subject is an infant or small child
  - Contrary to the common believe that subjects play no roles in selectional restriction
- The solid/liquid food contrast is a direct human experience after a child is weaned
- MARVS theory
  - *chi1/he1 nai3 shui3* subject-internal attributes
  - The Agent Role: [INGEST OBJECT [+/- Solid]]

Shakespearean-garden Approach to Specific Ontology: Tang Poems

唐詩三百首文本超過 150,000 字

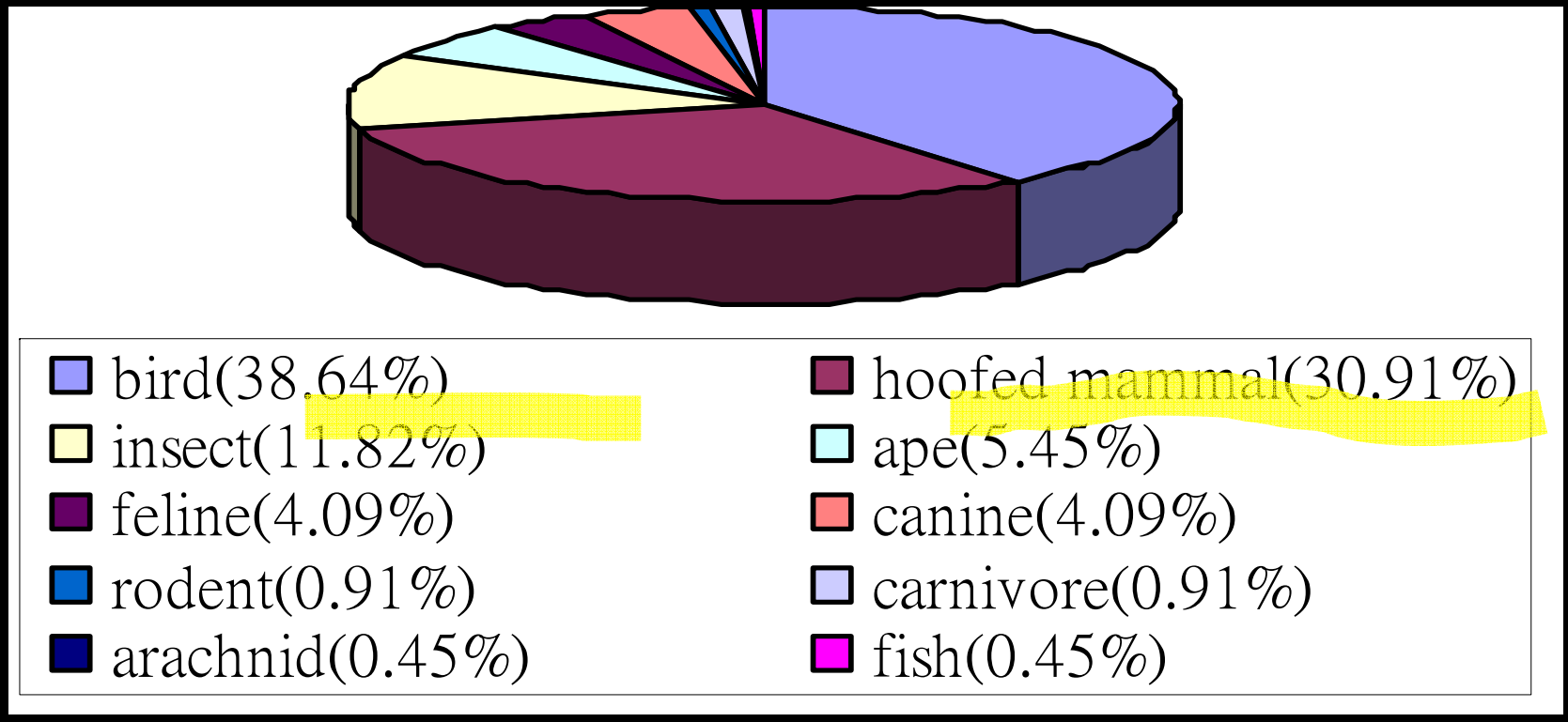
- 全部斷詞並分類
- 整理出三個不同領域的詞表（共計176個詞種）
  - 動物（64 個詞種）
  - 植物（59個詞種）
  - 人造物（53個詞種）

bow.sinica.edu.tw/ont/ts300_ont.html

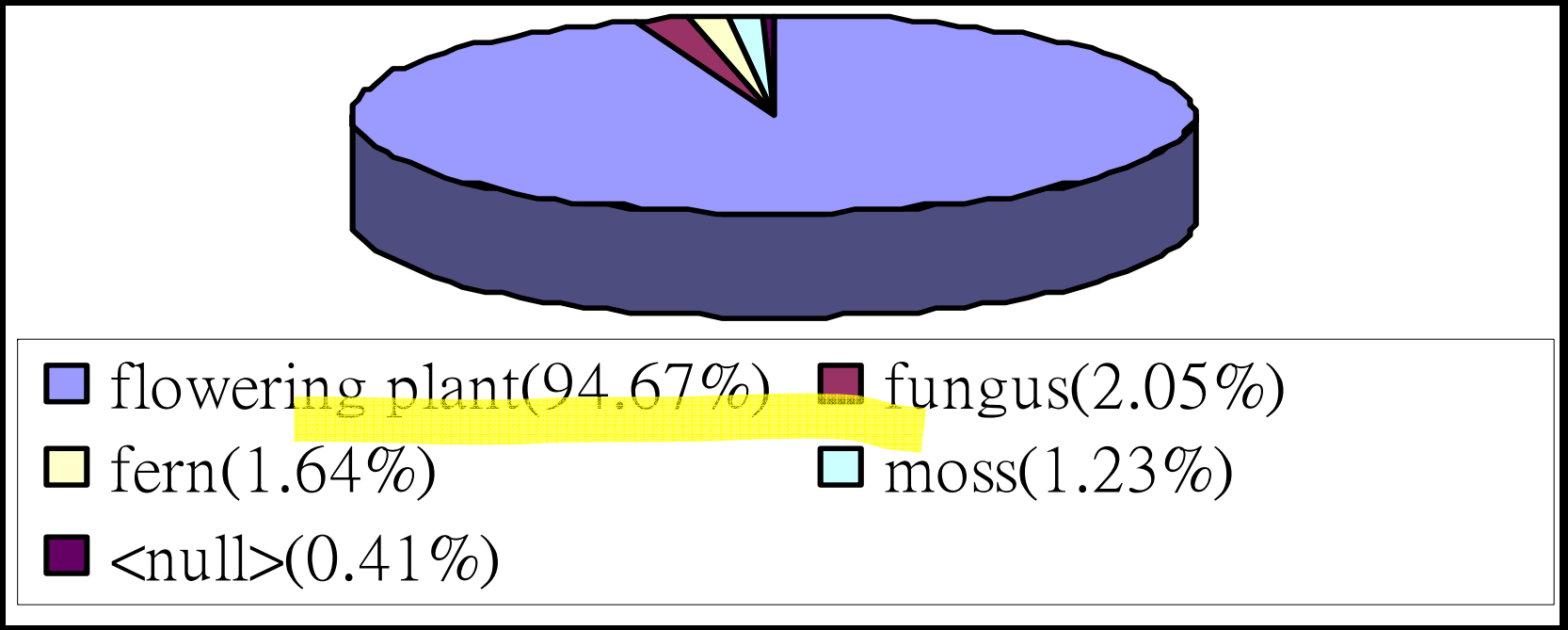
唐詩三百首領域詞彙庫示例-動物的知識架構

- 水棲哺乳動物 *No marine mammals
- 有蹄哺乳動物 馬 牛 羊 駱駝 斑騅 鹿
- 有袋類 *No marsupials
- 肉食性動物 羆 熊
  - 犬科動物 狼 豺 犬
  - 貓科動物 貙 lynx 貔 虎 tiger
- 嚙齒動物 鼯 蝙蝠
- 靈長類
  - 猿 猿 猩
  - 猴 猱 macaque

唐詩三百首中動物概念的分佈



唐詩三百首中植物概念的分佈



Knowledge Inferred from the Animal Ontology of Tang Poems

- 沒有有帶類動物（如袋鼠等）
  - 19世紀才在澳洲發現
- 沒看到海生哺乳動物/有眾多的偶蹄類動物
  - 唐文化以陸上活動爲主，並以馬爲重要活動工具
- 哺乳動物中鳥類佔及大部分/無脊椎動物中只有昆蟲，沒有其他類

  唐文化醉心於飛的意念（如：飛天）

  [無脊椎動物中只有昆蟲有翅膀。]

Conclusion I: Linguistic Insights

Corpus as Evidence

- Corpora greatly extend linguists' range of observations
  - Even across the barrier of time
- Language as an living organism allows variations and adaptations (the evolutionary view)
- The coherence of language is the shared tendency of all users
- Distributional data in corpus lead to discovery of these shared tendencies
  - This should be more valuable than incidental example

Conclusion II: Theoretical Relevance

Language as a Knowledge System

- The **generative lexicalist** approach to grammar: language as a knowledge system
- All aspects of Language are projected from a unified knowledge system
- Linguists need to make our theoretical account relevant to the world of knowledge and use, not vice versa
  - Failing to do so will get us/our theories deleted