WordNet Based Comparison of Language Variation : A study based on CCD and CWN

Jia-Fei Hong*, Chu-Ren Huang*, Yang Liu**

***Institute of Linguistics, Academia Sinica ,Taiwan**

****Institute of Computational Linguistics, Peking University**

GWC-06, Jeju Korea, Jan. 22-26, 2006

Outline

- Introduction
  - Goal
  - WordNet and CWN
  - WordNet and CCD
- Analysis of Translation Differences
  - Analysis by Synsets
  - Analysis by Translation Words
- Conclusion

Goal

- Compare the translations of WordNet English synsets in CCD and CWN
- CCD (Chinese Concept Dictionary)
  - China usage
  - Simplified Chinese characters
  - Peking University
- CWN (Chinese Wordnet)
  - Taiwan usage
  - Traditional Chinese characters
  - Academia Sinica

WordNet and CWN I

- **C**hinese **W**ord**n**et → **CWN**
- Developed by Academia Sinica Wordnet group
- A bilingual Chinese-English Wordnet, mapped to SUMO ontology (Niles & Pease, 2001)

WordNet and CWN II

- This research is using WN version 1.6
- All Wordnet 1.6 synsets (roughly100,000)
- For each English synset: 3 most appropriate Chinese translation equivalents (CTE)

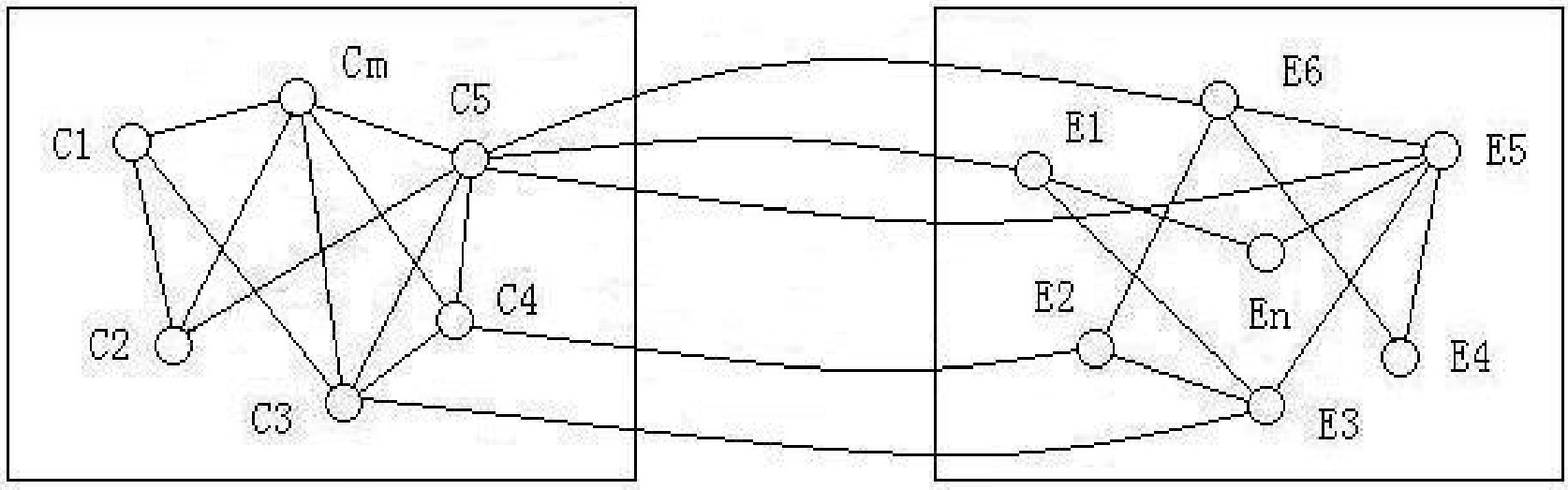
WordNet and CCD I

- **C**hinese **C**oncept **D**ictionary → **CCD**
- Developed by the Institute of Computational Linguistics, Peking University
- A bilingual Chinese-English Wordnet

WordNet and CCD II

- Maintain the original WordNet synsets
- Different descriptive structures between Chinese and English
- Focus of the structure in CCD
  - presentation of the concept defined by a synset

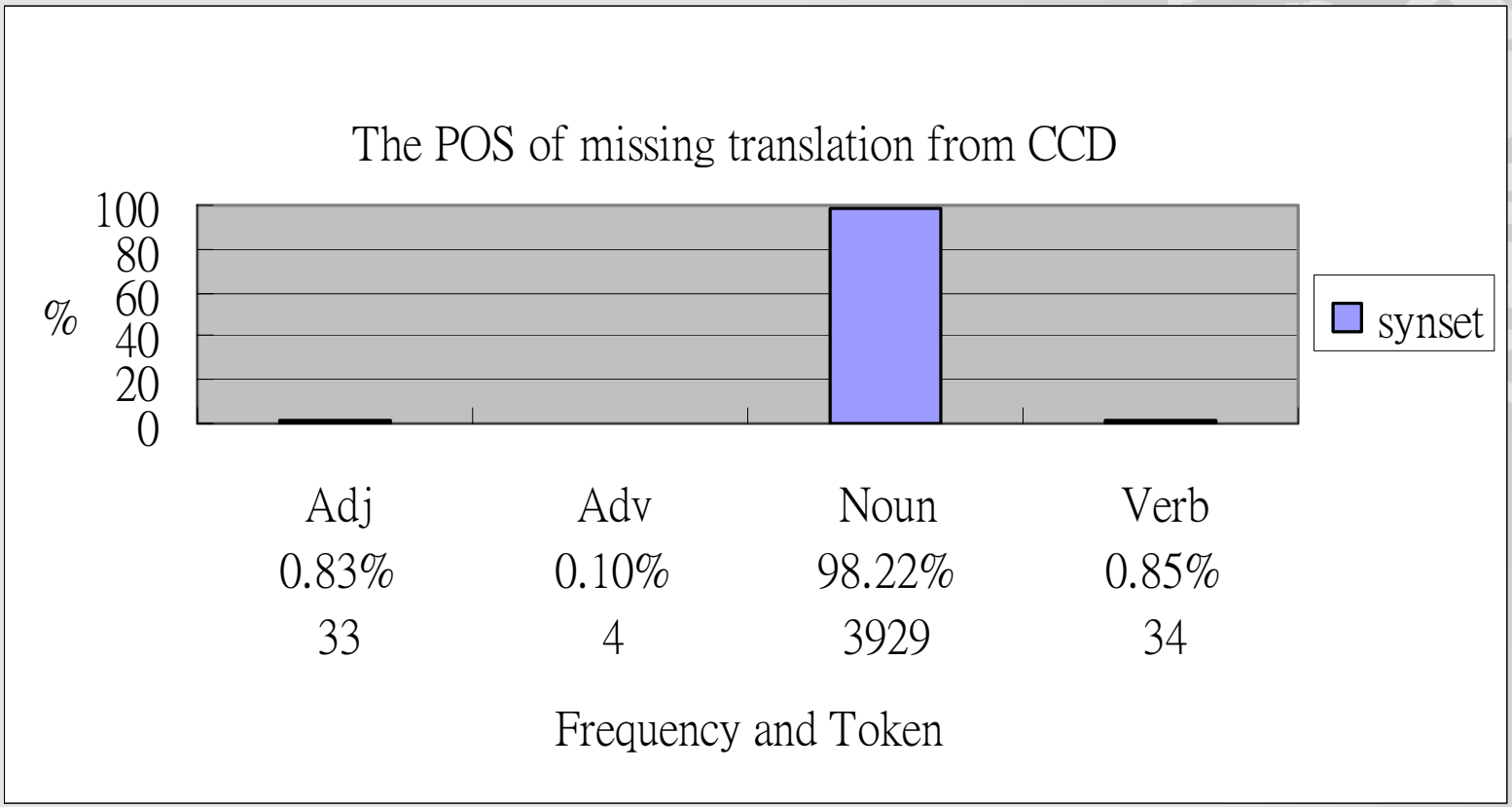
The Concept Relationships for Each Sub-network



Mis-matched Translation

- ALL English synsets are translated in CWN
- But some English synsets are not translated in CCD
  - Most of missing translations in CCD are nouns

Synsets Not Translated in CCD



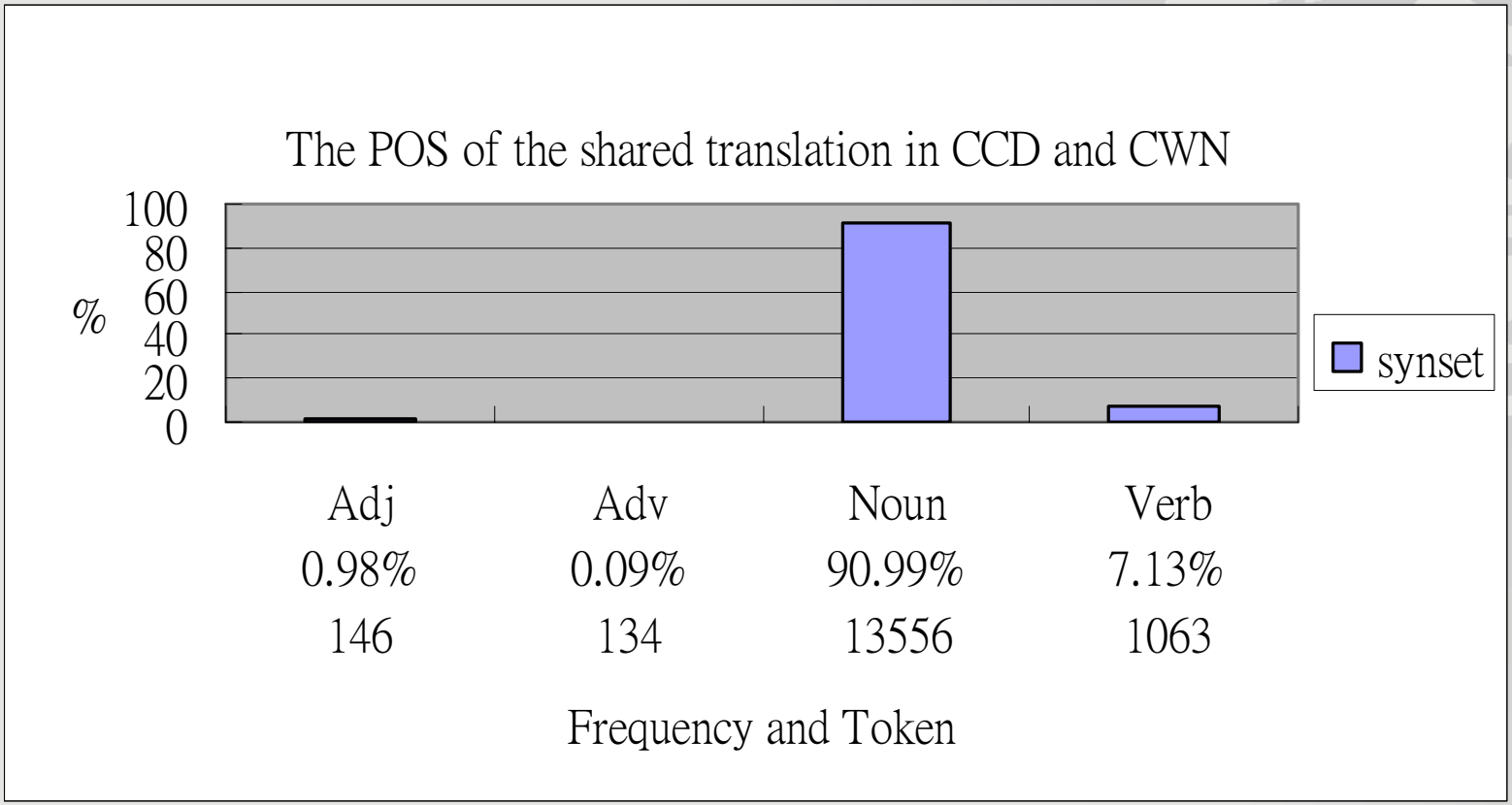
Synsets Not Translated in CCD: Examples

Meaning	Synset	CCD Translation	CWN Translation
colloquial British abbreviation for chocolate ice cream	choc-ice	X	巧克力冰淇淋 (qiao ke li bing qi lin)
provide with traffic signals, as of an intersection	signalize	X	向…發信號 (xiang…fa xin hao)

Translation Differences By Synsets

- In CCD and CWN translations, English synsets can have :
  - Shared translation: total 14,899 tokens
  - Unique translation in CCD: total 79,408 tokens (one term in China usage, but a phrase in Taiwan usage)
  - Unique translation in CWN : total 46,708 tokens (one term in Taiwan usage, no translation in CCD)

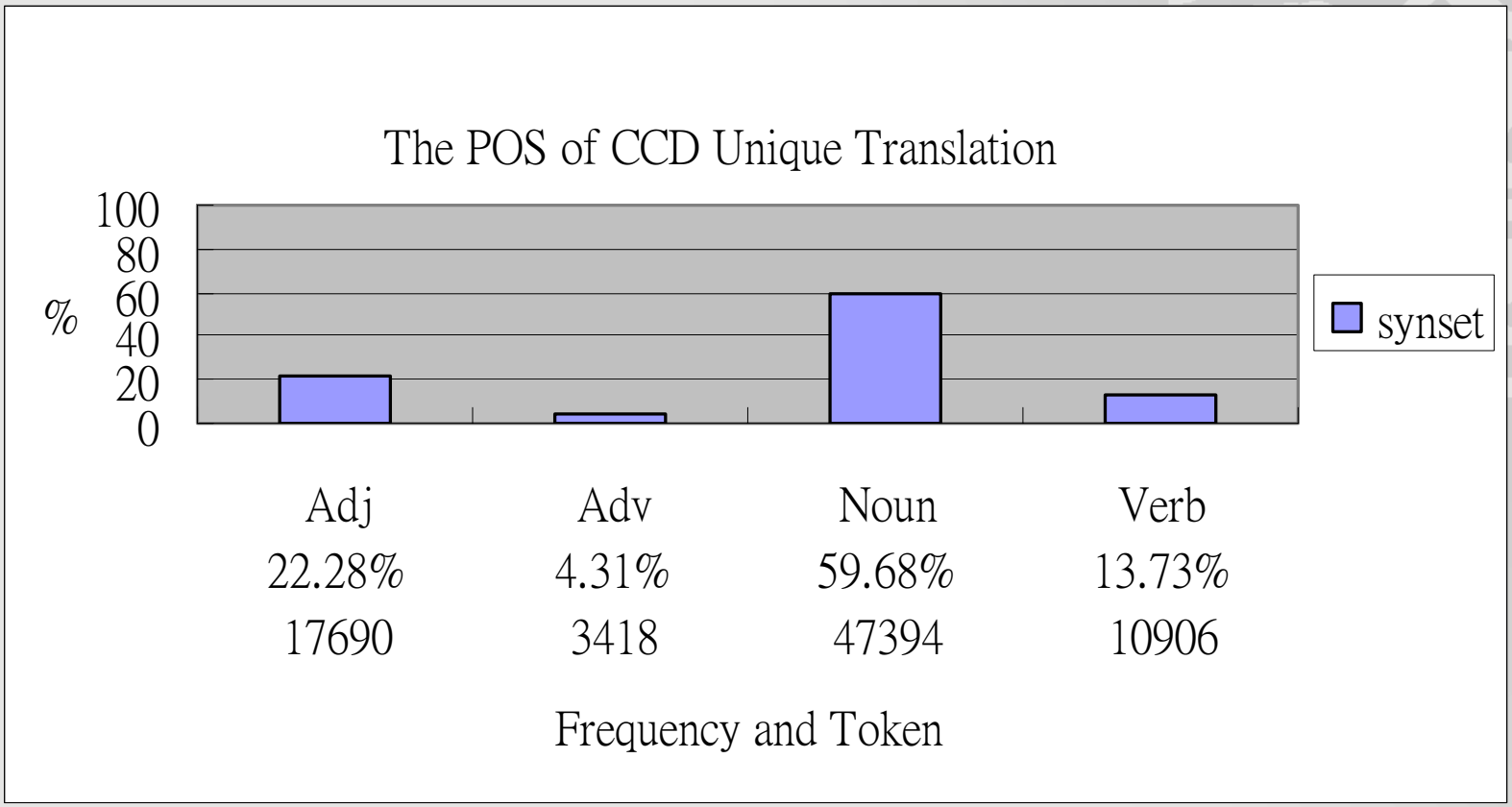
Shared Translations in CCD and CWN



Examples for Shared Translations

Meaning	Synset	Shared Terms
a shelf on which to keep books	bookshelf	書架(shu jia) 書櫃(shu gui) 書櫥(shu chu)
text that is typed or printed on paper	hard copy	硬複本(ying fu ben) 硬拷貝(ying kao bei) 硬式複本(ying shi fu ben) 硬性複本(ying xing fu ben)

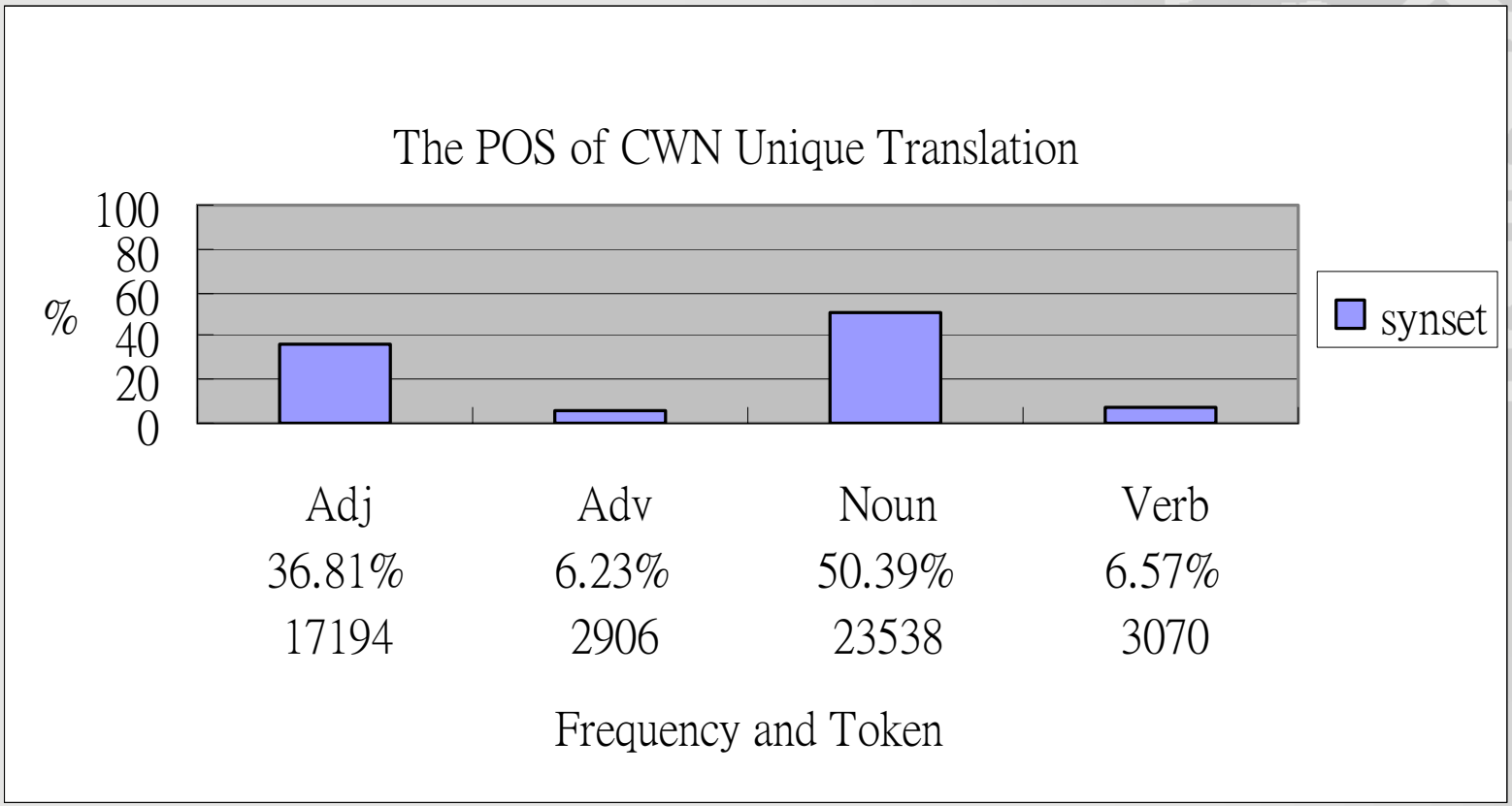
CCD Unique Translation



Examples for CCD Unique Translation

Meaning	Synset	CCD Unique Translation
a period of time assigned for work	hours	課時(ke shi)
the spreading of something (a belief or practice) into new regions	propagation, extension	外延(wai yan)

CWN Unique Translation



Examples for CWN Unique Translation

Meaning	Synset	CWN Unique Translation
a list of commodities that are not subject to tariffs	free list	免稅貨物單 (mian shui huo wu dan)
a computer small enough to use in your lap	laptop	筆記型電腦 (bi ji xing dian nao)

Using CCD Data to Improve CWN

- **Improvements are made when CCD Unique translation is adopted in CWN**

	Adjective	Adverb	Noun	Verb
Total data	17690	3418	47394	10906
Examinate	4000	3418	4000	4000
Improvement	215	76	118	128
Percentage	1.22%	**2.22%**	**0.25%**	1.17%
		highest	**lowest**	

Examples: CCD Improvements for CWN

Meaning	Synset	CCD Translation	CWN Translation
more often or more frequently	oftener	動輒(dong zhe) 經常(jing chang)	更經常地 (geng jing chang de) 頻率更大地 (pin lyu geng da de)
the act of discovering or expressing the quantity of something	quantification	量化(liang hua)	定量(ding liang)

Translation Differences in CCD and CWN by POS

The distribution of translation difference in CCD and CWN

	Adjective	Adverb	Noun	Verb	Total
Number of synsets	17915	3575	66025	12127	***99642***
Translation not shared	5023	1055	17265	4031	***27374***
	28.03%	**29.51%**	**26.15%**	**33.23%**	***27.47%***
			lowest	**highest**	

Examples of Translation Differences

- **masterstroke** (Noun) : an achievement demonstrating great skill or mastery
  - CCD：妙舉(miao ju)；CWN：神技(shen ji)
- **lay off** (Verb) : dismiss, usually for economic reasons
  - CCD：下崗(xia gang)；CWN：解雇(jie gu)

- Language Variation Between China and Taiwan, or
- Translator's mistake

Translation Difference by Translation Words

The distribution of Unique Translation words

Category	CCD only	CWN only	Shared	Total
Token	177,174	116,043	127,298	420,515
Percentage	42.13%	27.60%	30.27%	100%

The Situation for Adjective in Case of Translation Differences

- **Shared compound head:**
  - black: offering little or no hope
    - CCD：黯淡(an **dan**)； CWN：慘淡(can **dan**)
- **Shared compound element:**
  - according: (followed by to) as reported or stated by
    - CCD：據報(**ju** bao)； CWN：根據(gen **ju**)
- **No shared element:**
  - off: not performing or scheduled for duties
    - CCD：離崗(li gang)；CWN：休假的(xiou jia de)

Distribution of Adjective Semantic Relation

For Adjectives

Category	Shared compound head	Shared compound element	None	Total
Synset	861	94	5023	5978
	955			
Percentage	14.40%	1.58%	**84.02%**	100%
	15.98%			

Distribution of Adverb Semantic Relation

For Adverbs

Category	Shared compound head	Shared compound element	None	Total
Synset	477	51	1055	1583
	528			
Percentage	30.13%	3.22%	66.65%	100%
	33.35%			

Distribution of Noun Semantic Relation

For Nouns

Category	Shared compound head	Shared compound element	None	Total
Synset	11173	881	17265	29319
	12054			
Percentage	38.11%	3.00%	58.89%	100%
	41.11%			

Distribution of Verb Semantic Relation

For Verbs

Category	Shared compound head	Shared compound element	Difference	Total
Synset	5640	720	4031	10392
	6360			
Percentage	54.28%	6.93%	38.79%	100%
	61.21%			

Conclusion I:

GOAL

- Use two versions of English-Chinese Bilingual Wordnet to improve each other
- Show that WordNet can be a powerful tool for study of language variation

Conclusion II

- Analysis of results:
  - A big percentage of lexical variations between China and Taiwan are still anchored by shared concept, expressed by shared character in a compound word
  - Adjectives appear to be the most variable (nearly 85% of differences are not-related)
  - Verbs have the most conceptually consistent translations (Note. Verbs also have the most synonymous translations from English to Chinese in our study, Huang et al. 2006)

Acknowledge

- Thank all members of the CWN group.
- http://bow.sinica.edu.tw
- Thank the CCD research team.
- http://icl.pku.edu.cn

Thank YOU!