# A Bidirectional Study of Mandarin Conversation Verbs*

**Yiching Wu**
**d898702@oz.nthu.edu.tw**
**Institute of Linguistics National Tsing Hua University**

**Abstract**

This study examines verbs of conversation, from two directions, bottom-up and top-down, e.g. 交談 *jiao1 tan2* 'talk', 商量 *shang1 liang2* 'discuss', 吵架 *chao3 jai4* 'quarrel', and 聊天 *liao2 tian1* 'chat' etc. In addition to the inductive bottom-up method, inducing generalization on the semantic properties of a lexical item by identifying its syntactic behavior and collocations, the deductive top-down approach, deducing semantic attributes from domain ontology is found to be helpful in systematically accounting for the linguistic phenomena.

## 1. Introduction

There are two common strategies used to determine truth from facts, induction and deduction. Studying lexical semantics is no exception. Linguists also probe lexicons from bottom-up or top-down perspectives.

### 1.1 Bottom-up approach: from lexical items to semantic fields

By following this approach, linguists may study from either a single lexical item (e.g. Fillmore and Atkins 1992), a pair or a set of near synonyms (e.g. Tsai et al 1996, 1998, Chief et al 2000, Liu et al 2000, Wu and Liu 2001, Liu 2000, 2002a, 2002b, 2003, etc.), or a class of lexical items (Chang et al 2000b, Lien 2001 & 2002, etc.) in order to capture the generalization of semantic components, constraints and rules for a semantic field, thereby constructing their theories. Generalizations may be derived from an observation of syntactic behavior and collocations of the items. The linguistic data may be collected from linguists' own intuition, informants' judgment, dictionaries, or from electronic thesauri e.g. WordNet (http://www.cogsci.princeton.edu/~wn/index.shtml/), and corpora such as British National Corpus (BNC) at http://www.hcu.ox.ac.uk/BNC/, and Sinica Corpus at http://www.sinica.edu.tw/ftms-bin/kiwi.sh/.

### 1.2 Top-down approach: from upper classes to lexical items

Using this approach, linguists start from an upper class, probe their way through the subclasses, and then to specific lexical entries. In general, the aim of this method is to facilitate language processing by constructing a taxonomy or ontology of the human lexicon. Semantic hierarchy and inheritance relations are the two main research targets. HowNet (http://www.keenage.com/html/c_index.html/) and Suggested Upper Merged Ontology

* This study is supported in part by NSC 91-2411-H-009-012-ME. I am indebted to my colleagues in the project and two anonymous reviewers for their insightful comments.

(SUMO at http://ontology.teknowledge.com/) are two of the online representatives. They contain a nearly complete hierarchy for Chinese and English words respectively. VerbNet (http://www.cis.upenn.edu/verbnet/) based on Levin (1993) and FrameNet I (http://www.icsi.berkeley.edu/~framenet/) are two of the other less exhausted cases. In Levin (1993), there are forty-eight verb classes grouped by a variety of syntactic alternations, but these classes are not structured by other upper classes. Though the concept of domains is obliterated in FrameNet II, FrameNet I contains fourteen domains with subordinate frames and lemmas, but these domains are not subsumed to other superior classes.

The problem of the bottom-up approach is that the semantic properties of each lexical item may be extracted and the overt syntactic behavior may be accounted for, but the inheritance relationship, with its parent and ancestor classes, remains opaque. In contrast, the problem surrounding the top-down approach is that the inheritance relationship among the different levels may be clear enough to account for the covert syntactic behavior, but the detailed semantic attributes may be missed. To compensate for this drawback, SUMO combines its ontology with WordNet synsets. (Pease et al 2002), and researchers are now pursuing a multi-lingual semantic network (Huang et al 2002). A prototype of the Chinese-English bilingual interface of general and domain-specific ontologies, constructed by the Chinese Knowledge Information Processing Group (CKIP), is now also available at http://godel.iis.sinica.edu.tw/CKIP/ontology/.

This study aims to provide a bidirectional approach, incorporating the above two methods in order to explore a detailed analysis of the finer semantic distinctions of conversation verbs.

## 2 Conversation verbs

To extract Chinese conversation verbs, several resources were consulted. Firstly, Conversation is one of the fourteen frames of the Communication domain in FrameNet I, and there are both Chinese and English words, as well as definitions, in HowNet. By retrieving the corresponding Chinese words and definitions of the English lemmas subsumed to the Conversation frame in FrameNet, a set of possible Chinese candidates is obtained. Secondly, the resultant set of candidates was checked with the lexical items in CKIP's Chinese-English bilingual ontologies. Any items that are used only in mainland China were temporarily ruled out. Thirdly, dictionaries, thesauri, and the intuition of native speakers were consulted. Finally, entries and their frequency in Sinica Corpus were taken into consideration. In this way, a set of target Chinese conversation verbs was obtained, e.g. 交談 *jiao1 tan2*, 談話 *tan2 hua4,* 會談 *hui4 tan2* 'talk', 閒聊 *xian2 liao2* 'gab' and 聊天 *liao2 tian1* 'chat', 交流 *jiao1 liu2*, and 溝通 *gou1 tong1* 'communicate', 商量 *shang1 liang2*, 討論 *tao3 lun4*, and 商討 *shang1 tao3* 'discuss', 吵架 *chao3 jia4* 'quarrel' and 爭辯 *zheng1 bian4* 'debate', etc.

After setting the target items, their syntactic behavior and collocations were probed. In

addition, their upper class, the domain of communication was also investigated. In what follows, we will first illustrate how the near synonyms were analyzed from a bottom-up approach and then elaborate on a top-down method.

**3. Analysis of near synonyms**

In this section, we will use three verbs of 'talk/converse' as an example to illustrate the bottom-up approach: *jiao1 tan2*, *tan2 hua4* and *hui4 tan2*, literally meaning 'talk to each other', 'talk words' and 'meet and talk' respectively.

**3.1 Grammatical function distribution**

As shown in table 1 below, sixty percent of the *jiao1 tan2* tokens function as a predicate, the main verb of a clause. In contrast, the majority of the tokens of *tan2 hua4* and *hui4 tan2* are used as a head noun. The three lexical items have approximately the same functional percentage as a modifier.

| Lemma / Function | 交談 *jiao1 tan2* | 談話 *tan2 hua4* | 會談 *hui4 tan2* |
|---|---|---|---|
| Predicate | 71 (60%) | 57 (31%) | 50 (33%) |
| Head Noun | 30 (25%) | 101 (54%) | 82 (54%) |
| Modifier | 17 (15%) | 28 (15%) | 20 (13%) |
| Total | 118 (100%) | 186 (100%) | 152 (100%) |

Table 1: Grammatical function distribution of *jao1 tan2*, *tan2 hua4*, and *hui4 tan2*

**3.2 Collocation**

All three verbs can be modified by a duration, e.g. *Ta1 men jiao1 tan2/hui4 tan2 le shi2 fen1 zhong1* and *Ta1 men tan2 le shi2 fen1 zhong1 de hua4* 'They have talked for ten minutes'. The three verbs can all collocate with the progressive (正 *zheng4*) 在 *zai4* and the experiential 過 *guo4*, e.g. *Ta1 men zheng4 zai4 jiao1 tan2/tan2 hua4/ hui4 tan2* 'They are talking to each other,' and *Ta1 men jiao1 tan2 guo4 /tan2 guo4 hua4/hui4 tan2 guo4* 'They have talked to each other.' In addition, they can all be followed by the inchoative particle 了 *le*, e.g. *Ta1 men* (*kai1 shi3*) *jiao1 tan2/tan2 hua4/hui4 tan2 le!* 'They start to talk!' From the above facts, and by following the methodology used by Chang et al (2000a), we can induce the generalization that these verbs are bounded process verbs. However, these verbs contrast with 'discuss' verbs such as 商量 *shang1 liang2* and 討論 *tao3 lun4* in that they do not take a Topic directly, e.g. '**jiao1 tan2/*tan2 hua4/*hui4 tan2/shang1 liang2/tao3 lun4 shi4 qing2* '*converse/discuss about something'. Furthermore, they do not take a Message in the same manner as other saying verbs, e.g. *'Ta1 men *jiao1 tan2/*tan2 hua4/*hui4 tan2/shuo1 ta1 men mei2 you3 qian2 'They *converse*d/*talked/said they had no money.'*

In addition, the subject agent, the Speaker, of the three verbs must be plural, e.g. *Ta1*

*gen1 wo3 /wo3 men/ *wo jiao1 tan2 le ban4 xiao3 shi2/tan2 le ban4 xiao3 shi2 de hua4/hui4 tan2 le ban4 xiao3 shi2* 'He and I /we/*I have talked for half an hour.' This symbolizes the reciprocality of a conversation event, in which both the speaker and the listener do the speaking and listening. However, *hui4 tan2* differs from the other two in that its speakers are mostly officials. When *hui4 tan2* functions as a predicate, only 18% (9/50) of the Speakers are common people. Most Speakers (82%) are government officials, representatives of countries or parties, or school officials. In addition, among the nine instances of non-officials there are two doctor-and-patient pairs, and two businessmen pairs.

When the Speakers are realized as Interlocutor_1 and Interlocutor_2, being an argument in a matrix clause or in a subordinate clause as a pre-nominal modifier, they may be linked with or without an overt connective such as 與*yu3*, 和*he2/han4*, and 跟*gen1*, e.g. 戈巴契夫與葉爾辛/美國國務卿貝克和伊拉克總統海珊/我們跟所有相關的人士/辜汪 *ge1 ba1 qi4 fu1 yu3 ye4 er3 xin1/mei3 guo2 guo2 wu4 qing1 bei4 ke4 he2/han4 yi1 la1 ke4 zong3 tong3 hai3 shan1/wo3 men gen1 suo3 you3 xiang1 guan1 de ren2 shi4/gu1 wang1* 'Gorbacheve and Yeltsin/the American Secretary of State, James Baker, and the President of Iraq, Saddam Hussein/we and all the related people/Koo and Wang'. Among these three overt connectives and the covert linker, *gen1* is the most colloquial and is often used in daily conversation, whereas *yu3* and the covert linker usually appear in formal texts**.** There are seventy-one instances of Interlocutor_1 and Interlocuteor_2 using *hui4 tan2* in Sinica Corpus. The distribution of the four linking devices is shown in table 2 below.

| pattern / count | Interlocutor_1 **conj.** Interlocutor_2 | | | Interlocutor_1 Interlocutor_2 |
|---|---|---|---|---|
| | 與 *yu3* | 和 *he2/han4* | 跟 *gen1* | covert linker |
| total | **46** | 9 | 1 | **15** |

Table 2: Linking devices

*Yu3* and the covert linker connect forty-six and fifteen pairs of speakers respectively. 和 *he2/han4* links nine, but 跟 *gen1* combines only one. This shows that *hui4 tan2* is a formal conversation event.

### 3.3 Lexical Distinctions Redefined as the MARVS Representation

The above generalizations can be represented by the Module-Attribute Representation of Verbal Semantics (MARVS) proposed by Huang and Ahrens (1999) and Huang et al (2000).

| Module/Attributes | 交談 *jiao1 tan2* | 談話 *tan2 hua4* | 會談 *hui4 tan2* |
|---|---|---|---|
| Event Module<br>Inherent Attributes | ●//////●<br>\|<br>[Reciprocal] | ●//////●<br>\|<br>[Reciprocal] | ●//////●<br>\|<br>[Reciprocal] [formal] |
| Role Module<br>Role-Internal Attributes | ＜Speaker, Medium＞<br>\| \|<br>[Plural] [language] | ＜Speaker＞<br>\|<br>[Plural] | ＜Speaker＞<br>\|<br>[plural][representative ] |

Fig. 1: MARVS Representation of the semantic differences among conversing verbs

From the above discussion, we can induce the following generalizations. Firstly, each of the above three items denotes a bounded process event which refers to a reciprocal communication activity. Since it is a reciprocal event, the Speaker role must have a minimum of two agents. Secondly, *hui4 tan2* is a more formal conversational event in contrast with the other two, and thus its Speakers tend to be representatives of a country or an organization. Thirdly, *jiao1 tan2* is inclined to take a language Medium whereas *tan2 hua4* and *hui4 tan2* do not. In addition, we know that the 'talk/converse' verbs do not collocate with a Topic as with the 'discuss' verbs, nor do they co-occur with a Message as with the 'say' verbs. However, as we cannot adequately account for them so far, we will attempt an alternate approach in the next section.

**4. From a domain, frames, to subframes**

In this section, we will take a top-down perspective to investigate the verbs of conversation. In FrameNet, there are fourteen frames within the domain of Communication. To capture the conceptual structure for understanding events in the domain of communication, Liu and Wu (2003) propose a schematic representation as shown in Fig. 2 below:

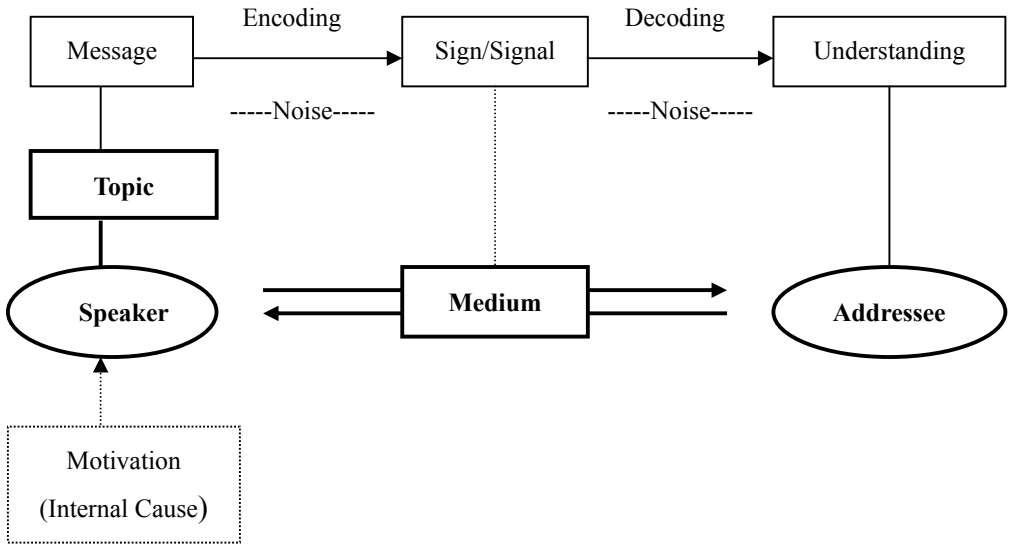



Fig. 2: Schematic Representation of Conversation

Communication in general is realized as an information-exchange process, where a Speaker, from certain motivation, sends a Message on a given Topic, through a process of packaging (Encoding), and an Addressee receives the package, decodes it, and reaches a certain understanding. The process is reciprocal and is carried out via a Medium (face-to-face, phone, TV, or email and fax, etc.).

Speaker, Addressee, Topic, Message, Sign/Signal, and Medium are the core frame

elements (FEs) of Communication. Each of the fourteen frames of Communication profiles certain frame elements. The Conversation frame focuses on the bilateral communication between the Speaker and the Addressee which are realized as Interlocutor_1, Interlocutor_2 and Interlocutors. Therefore, in addition to the three roles, only Medium and Topic are possible participant roles. The above schema may also account for the reason why Topic is not obligatory to all Chinese verbs, and Message is not a core element in the Conversation frame. Since the central focus is on the reciprocal communication process, Topic may not be profiled in every case, and Message may be suppressed.

Conversation verbs may be further classified into four subtypes according to their different purposes and manners:

| Subframe | Purpose | Manner | Highlighted FEs |
|---|---|---|---|
| 1 Converse | to exchange information | unmarked | Medium-language |
| 2 Discuss | to solve a problem | serious | Topic |
| 3 Quarrel | to exchange different opinions | heated | Cause |
| 4 Chat | for fun | causal | Accompanying activities |

Table 3: Subframes of Conversation

The Converse subframe is unmarked with a purpose to exchange information, e.g. *jiao1 tan2*, *hui4 tan2* 'talk', 交流 *jiao1 liu2* and 溝通 *gou1 tong1* 'communicate', etc. Hence, the Converse subframe verbs tend to co-occur with a language medium, e.g. *yi3 he2 lan2 hua4 jiao1 tan2* 'converse in Dutch'. In the Discuss subframe, interlocutors communicate in a more serious manner in order to solve problems, e.g. 商量 *shang1 liang2*, 討論 *tao3 lun4* and 商討 *shang1 tao3* 'discuss', therefore the verbs tend to collocate with a Topic, e.g. *shang1 liang2 jie2 hun1 de shi4* 'discuss a wedding affair' and *tao3 lun4 nong2 ye4 wen4 ti2* 'discuss issues on agriculture'. In the Quarrel subframe, interlocutors exchange different opinions in a heated manner, e.g. 吵架 *chao3 jia4* 'quarrel' and 爭辯 *zheng1 bian4* 'debate'. Verbs in this subframe tend to collocate with a cause that results in the disagreement, e.g. *wei4 le qian2 chao3 jia4* 'quarrel about money'. In the Chat subframe, interlocutors communicate in a casual manner for fun, e.g. 閒聊 *xian2 liao2* 'gab' and 聊天 *liao2 tian1* 'chat', etc., and hence the verbs tend to co-occur with accompanying recreational activities such as drinking coffee, e.g. *he1 ka1 fei1 liao2 tian1* 'drink coffee and chat'.

From this point of view, the collocation of a Topic with 'discuss' verbs, as well as other highlighted participant roles in the subframes, may also be systematically accounted for.

## 5. Conclusion

The conversation verbs studied here serve to illustrate a hybrid approach to lexical semantics. The bottom-up approach provides a detailed generalization from studying specific lexical items. The top-down approach, aided by the domain schema, provides an overall outlook of the properties of the whole domain, helping to offer a systematic account for the linguistic phenomena. Although each of the methods has both positive and negative aspects, by incorporating the two approaches, detailed semantic features and outlined semantic properties can be expected.

## References


Chang, Li-Li, Keh-Jiann Chen and Chu-Ren Huang. 2000a. A Lexical-Semantic Analysis of Mandarin Chinese Verbs: Representation and Methodology. *International Journal of Computational Linguistics and Chinese Language Processing*. 5(1).1-18.

Chang, Li-Li, Keh-Jiann Chen and Chu-Ren Huang. 2000b. Alternation Across Semantic Field: A Study of Mandarin Verbs of Emotion. *International Journal of Computational Linguistics and Chinese Language Processing*. 5(1).61-80.

Chief, Lian-Cheng, Chun-Ren Huang, Keh-Jiann Chen, Mei-Chih Tsai, and Li-Li Chang. 2000. What Can Near Synonyms Tell Us? *International Journal of Computational Linguistics & Chinese Language Proceeding*. 5.1, 47-60.

Fillmore, Charles J., and Atkins, Beryl T. 1992. Toward a Frame-Based Lexicon: The Semantics of RISK and Its Neighbors. *Frames, Fields, and Contrasts*, ed. by Adrienne Lehrer and Eva Feder Kittay. 75-102. Hillsdale. New Jersy: Lawrence.

Huang, Chu-Ren, and Kathleen Ahrens. 1999. The Module-Attribute Representation of Verbal Semantics. *Working Papers on Chinese Verbal Semantics 1*, ed. by Kathleen Ahrens, Chu-Ren Huang, and Mei-Chih Tsai. 1-14. Taipei: Academia Sinica.

Huang, Chu-Ren, I-Ju E. Tseng, and Dylan Tsai. 2002. "Translating Lexical Semantic Relations: The First Step Towards Multilingual Wordnets," presented at SemaNet'02: Building and Using Semantic Networks, A COLING2002 Post-Conference Workshop. Aug. 31. Academia Sinica. Taipei.

Huang, Chu-Ren, Kathleen Athens, Li-Li Chang, Keh-Jiann Chen, Mei-Chun Liu, and Mei-Chih Tsai. 2000. The Module-Attribute Representation of Verbal Semantics: From Semantics to Argument Structure. *International Journal of Computational Linguistics and Chinese Language Processing*. 5(1).19-46. Also appeared in *Proceedings of the Symposium on Selected NSC Projects in General Linguistics from 1998-2000.* 119-46. 2001.

Lien, Chinfa. 2001. Verbs of Saying in Li Jing Ji. The 4th Interenational Conference on Classical Chinese Grammar. University of Britishi Columbia, 15-17 August.

Lien, Chinfa. 2002. Lexicalization and Grammaticalization in Taiwan Southern Min—A Case

Study of Verbs of Commercial Transaction. To appear in *Joy of Linguistics*. Taipei: The Crane Publishing Co., Ltd.

Liu, Mei-Chun, Chu-Ren Huang, Charles Lee, and Ching-Yi Lee. 2000. When Endpoint Meets Endpoint: A Corpus-based Lexical Semantic Study of Mandarin Verbs of Throwing. *International Journal of Computational Linguistics & Chinese Language Proceeding*. 5.1, 81-96.

Liu, Mei-Chun. 2000. Categorical Structure and Semantic Representation: Mandarin Verbs of Communication. Paper presented at the 5th Conference on Conceptual Structure, Discourse and Language. University of California, Santa Barbara.

Liu, Mei-Chun. 2002a. Corpus-based Lexical Semantic Study of Verbs of Doubt: Huayi and Cai in Mandarin. Concentric. 28.2.

Liu, Mei-Chun. 2002b. Mandarin Verbal Semantics: A Corpus-based Approach. 2nd ed. Taipei: Crane.

Liu, Mei-Chun. 2003. From Collocation to Event Information: the Case of Mandarin Verbs of Discussion. To appear in *Language and Linguistics*.

Lui, Mei-Chun and Wu, Yiching. 2003. Beyond Frame Semantics: Insight from Mandarin Verbs of Communication. Paper presented at the 4th Chinese Lexical Semantics Workshop. (第四屆漢語詞彙語義學研討會), Department of Chinese, Translation, and Linguistices, City University of Hong Kong, Hong Kong. June 22-July 11. (http://icl.cityu.edu.hk/conference/4CLSW/BIG5/home.htm）

Niles, I., and Pease, A. 2001. Origins of the Standard Upper Merged Ontology: A Proposal for the IEEE Standard Upper Ontology. In Working Notes of the IJCAI-2001 Workshop on the IEEE Standard Upper Ontology, Seattle, Washington, August 6, 2001.

Pease, A., Niles, I., and Li, J. 2002. The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. In Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web, Edmonton, Canada, July 28-August 1, 2002.

Tsai, Mei-Chih, Chun-Ren Huang, Keh-Jiann Chen, Kathleen Ahrens. 1998. Towards a Representation of Verbal Semantic: An Approach Based on Near-Synonyms. *International Journal of Computational Linguistics & Chinese Language Proceeding*. 3.1, 62-74.

Tsai, Mei-Chih, Chun-Ren Huang, Keh-Jiann Chen. 1996. You Jinyici Bianyi Biaozhun Kan Yuyi, Jufa Zhi Hudong (由近義詞標準看語義、句法之互動 From near-synonyms to the interaction between syntax and semantics), paper presented at IsCLL-5, Taipei, Taiwan.

Wu, Yi-Ching, and Liu, Mei-Chun. 2001. The Semantic Distinction and Information Representation of Psychological Verbs: Xiang, Renwei, Yiwei and Juede—a Corpus Based Analysis. *Proceedings of Research on Computational Linguistics Conference XIV*. 317-336. Tainan: National Cheng Kung University.

# A Corpus-Based Study on Mapping Principles of Metaphors in Politics

Shu-Ping Gong

Graduate Institute of Linguistics

National Taiwan University

d91142001@ntu.edu.tw

## Abstract

This study proposes a corpus-based method to generate Mapping Principle of metaphors. In particular, Ahrens's (2002) Mapping Principle in the Conceptual Mapping Model (CM model) is simply based on the native speakers' intuition instead of analyzing it from huge linguistic data. In order to provide more convincing evidence to support the CM model, we adopt the corpus method to extract out the metaphorical expressions in politics from the Academic Sinica Balanced Corpus. We analyze the correspondences existing within the source-target domain pairings and generate Mapping Principle based on the salient meanings in these linguistic expressions. We adopt this method to examine the mapping principles of five metaphors: POLITICS IS BUILDING, POLITICS IS A JOURNEY, POLITICS IS A PLAY, POLITICS IS A COMPETITION and POLITICS IS SPORT. This corpus-based method can provide a more convincing way to generate Mapping Principle at the linguistic level than the original one (Ahrens 2002).

**1. Introduction**

Lakoff and Johnson (1980) have proposed that metaphors are understood through the mapping from the concrete domain, i.e. the source domain, to the abstract domain, i.e. the target domain. For instance, the linguistic expression "You are wasting my time" is understood via mapping the MONEY domain to the TIME domain. However, the mapping principles between domains are not clearly defined. They do not explain how the mapping principles are generated and what constraints are governed.

Ahrens (2002) has proposed Conceptual Mapping Model to supplement the limit of Lakoff (1993) Contemporary Theory of Metaphors. She analyzes the lexical correspondences existing between a source and target domain in terms of "entities", "qualities" and "functions". Within this model, the underlying reasons for mapping can be generated based on the real linguistic data, so-called Mapping Principles.

One of the advantages is that her model can clearly point out that the inference exists between the source and target domains (Ahrens 2002: 275). The mapping principles are generated at the linguistic level by analyzing the expressions of conventional metaphors in a systematic method. In particular, she collects the metaphor examples from the native speakers' intuition, groups the metaphors into the source domains, analyzes them in terms of entity, quality and function based on the real world knowledge, and finally analyzes the mapping principles between the source and target domains. For instance, this model generates the mapping principle for IDEA IS BUILDING as "idea is understood as building because buildings involve a physical structure and ideas involve an abstract structure".

Another advantage is that this model proposes a Mapping Principle Constraint to explain why a target domain selects different source domains (Ahrens 2002: 279). The constraint proposes that a target domain will select only source domains that involve unique mapping principles. For example, IDEA can be mapped to the following domains, BUILDING, FOOD, COMMODITY, and INFANT. IDEA has different reasons to select these four source domains. IDEA selects the source domain of BUILDING to borrow the notion of structure; IDEA uses FOOD as the source domain to borrow conceptualization of intake and digestion; it chooses the source domain of COMMODITY to express the concepts of value and marketing; it uses the source domain of INFANT to borrow the concept of birthing process.

However, there is a problem for the method to generate the Mapping Principle in this model. The Conceptual Mapping Model (CM Model) lacks evidence to support whether Mapping Principle can really reflect the correspondences existing between the source and target domains. Even though the linguistic data this model uses for generating a Mapping Principle are collected from native speakers' intuition, the method how to collect the data

is not very clear. In particular, are the data collected from questionnaires or from interviews? How many speakers participated in the production task? In order to revise the weakness in the method applied by the Mapping Principle, Ahrens, Chung and Huang (2003) and Chung, Ahrens, and Huang (2003) have proposed a corpora-based operational definition for Mapping Principle. They examine 2000 random examples of "economy" (*jingji*) and generate Mapping principle based frequency in corpus. In addition, they integrate their Conceptual Mapping Model with SUMO to restrict the mapping principles.

This study follows the same methodology of Ahrens et al (2003) to examine *zhengzhi* "politics" in Mandarin. In particular, this study uses a quantitative method to explore the Mapping Principles. Instead of the linguistic expressions generated from native speakers' intuition, we use a corpus approach to collect linguistic materials. We analyze the mapping principles between the target domain POLITICS and the different source domains by extracting the metaphorical expressions in politics from the Academic Sinica Balanced Corpus, which can provide huge data of real language usages. In addition, based on the quantitative information, we can observe what are the salient and frequent mapping principles and determine what the potential mapping principle are generated. Section two will discuss how we adopt the corpus-based method to investigate the mapping principles of metaphors in politics.

**2. Corpora Data**

We collect data from Academic Sinica Balanced Corpus (1995), a tagged corpus of over 5 million words of modern Mandarin usage in Taiwan (http://www.sinica.edu.tw/SinicaCorpus/). First, we use *zheng4zhi4* "politics" as the searching keyword, and get the 1964 pieces of sentences containing the word *zheng4zhi4* "politics" from the Academic Sinica Balanced Corpus. Following the five steps of Mapping Principle analysis (Ahrens 2002), we extract 142 pieces of metaphorical expressions out of 1964 sentences and then categorize them into nine source domains: BUILDING, BUSINESS, COMPETITION, JOURNEY, OCEAN, PLAY, INVESTMENT, WAR and WEATHER. Only the source-target domain pairings which has more than ten instances are examined in this paper. In this study, we focus on the five metaphors: POLITICS IS BUILDING, POLITICS IS A JOURNEY, POLITICS IS A PLAY, POLITICS IS A COMPETITION, and POLITICS IS SPORT. In Tables 1-5 we show the numbers of sentences of each metaphor, the number of lexical tokens in the source domain for each metaphor in terms of "entity", "quality" and "function", as well as the postulated mapping principles.

For POLITICS IS BUILDING, we can see that all these correspondences between the source domain BUILDING and target domain POLITICS (Table 1) are related to the concept of "structure". A "structure" of a

building should associate with a base/foundation, a stable structure and formation. "Politics" uses the source domain "building" to conceptualize the notion "structure". A building doesn't fall down since it has a good foundation and a well-built/stable structure. Likewise, politics develops well if it has good structure and foundation.

It is worth noting that major of mappings (16 instances out of 19 ones) in the metaphor "politics is building" are related to "entities" of building, instead of "qualities" and "functions". Lakoff and Johnson (1980) mentioned the "ontological metaphors are ways of viewing events, activities, emotions, ideas etc., as entities and substances. Ontological metaphors serve various purposes, and the various kinds of metaphors there are reflect the kinds of purposes served." (Lakoff and Johnson 1980: 25-26). Thus, ontological metaphors can refer to the entity, qualify it, identify a particular aspect of it, see it as a cause, and act with respect to it. However, from the frequencies shown in Table 1, we can observe that all purposes ontological metaphors serve are not equally distributive. In the case, the metaphor POLITICS IS BUILDING puts emphasis on "referring to the structures/ model of building", instead of qualifying the stability of politics, or identifying the concept how to construct politics. Thus, we can generate the mapping principle as (1).

Table 1: POLITICS IS BUILDING (19 instances)

| | Metaphor | Frequency |
|---|---|---|
| Entities | *chu2xing2* "a small model" | 1 |
| | *ji1chu3* "base/foundation" | 1 |
| | *jie2gou4* "structure" | 11 |
| | *gou4tu2* "composition" | 2 |
| Qualities | *wen3ding4* "stable" | 1 |
| Functions | *jian4gou4* "to establish" | 1 |
| | *xing2cheng2* "to form | 2 |

**(1) Mapping principle for POLITICS IS BUILDING**

Politics is understood as building because building involves a physical structure and politics involve an abstract structure.

For POLITICS IS A JOURNEY, we can observe that all correspondences between the source and target domains (Table 2) are related to the concept of "traveling through roads/routes". In other words, the trip has starting and ending points; the travelers can stride or retreat on the route; they may encounter obstacles on the way of their trip; the roads can be bumpy. Likewise, "politics" can be conceptualized as a journey because the career of a politician has starting and ending points or because the road to democracy should be bumpy.

The metaphor POLITICS IS A JOURNEY, different from POLITICS IS BUILDING, puts emphasis both on entity (6 instances out of 17 ones) and function (10 instances out of 17 ones). In this case, viewing politics as a journey allow people not only to refer politics to "crossroad", "milestones", "target", "obstacles", etc., but also to point out (motivating) actions "retreat" and "stride" in politics, which are associated with events took place through traveling. Thus, we can generate the mapping principle as (2).

Table 2: POLITICS IS A JOURNEY (17 instances)

| | Metaphor | Frequency |
|---|---|---|
| Entities | *shi2zi4lu4kou3* "a crossroad" | 1 |
| | *mu4biao1* "target" | 1 |
| | *li3cheng2bei1* "milestone" | 1 |
| | *zhang4ai4* "obstacles" | 1 |
| | *zou3xiang4* "trend" | 1 |
| | *dao4lu4* "road" | 1 |
| Qualities | *kan3ke1* "bumpy" | 1 |
| Functions | *zou3shang4* "go up" | 2 |
| | *bu4ru4* "go into" | 1 |
| | *dao4tui4* "retreat" | 3 |
| | *mai4xiang4* " stride" | 2 |
| | *qi4bu4* "start to walk" | 1 |
| | *shang4gui3dao4* " "on the track" | 1 |

**(2) Mapping Principle for POLITICS IS A JOURNEY**

Politics is understood as a journey because a journey takes a traveler through physical roads/routes and politics takes a party/country/politician through abstract routes.

For POLITICS IS A PLAY, we can observe that all the correspondences between the source and target domains (Table 3) are related to "performance to the public". A play must have players, scripts and platforms. The purpose of a play is to provide performance/shows to entertain audience. Politics can be conceptualized as a play because in politics, politicians as players perform political shows to entertain (serve) their voters/citizens.

Table 3: POLITICS IS A PLAY (17 instances)

| | Metaphor | Frequency |
|---|---|---|
| Entities | *wu3tai2* "platform" | 10 |
| | *zheng4zhi4 xiu4* "political show" | 1 |
| | *yao4jiao3* "leading character" | 1 |
| | *jue2se4* "role" | 4 |
| Functions | *biao3yan3* "performance" | 1 |

All frequencies of mappings are related to the entities in play (16 instances out of 17 ones), suggesting that

the metaphor POLITICS IS A PLAY allows people to conceive politics as a play and make people to refer to "platform" and "role" in play. People are interest in what a role a politician in the political platform. This case does not put emphasis on qualifying politics as a tragedy and comedy or identifying the entertainment functions a play can provide. Thus, we can generate the mapping principle as (3).

**(3) Mapping Principle for POLITICS IS A PLAY**

Politics is understood as a play because a play involves players' performance on the platform to the audience and politics involves politicians' performance to the public.

For POLITICS IS A COMPETITION, we can observe that the correspondences in the source-target domain pairings (Table 4) are associated with the concepts of "conflict" and "competition". "Politics" is treated as a competition which is full of fights and conflicts.

Table 4: POLITICS IS A COMPETITION (18 instances)

| | Metaphor | Frequency |
|---|---|---|
| Qualities | *zheng1dou4* "conflict" | 1 |
| | *dou4zheng1* "conflict" | 8 |
| Functions | *jing4zheng1* "compete" | 5 |
| | *dou4zheng1* "conflict" | 4 |

The distributions of frequencies show that both qualifies and functions for a competition (both 9 instances out of 18 ones) are emphasized to describe politics. The proportion of mappings suggests that a competition is mapped to politics via qualifying to the property "conflict" and identify to the aspect "competing" involving in politics. Thus, we can postulate the Mapping Principle as (4),

**(4) Mapping Principle for POLITICS IS A COMPETITION**

Politics is understood as a competition because a competition is full of physical conflicts and politics is full of political/abstract conflicts.

For POLITICS IS SPORT, we can see that the correspondences in the source-target domain pairings are associated with the notion of "exercising". "Sport" involves physical exercise. "Politics" borrow this notion for being conceptualized as "mental exercising".

Table 5: POLITICS IS SPORT (16 instances)

| | Metaphor | Frequency |
|---|---|---|
| Entities | *jue3li4chang3* "a wrestling ring" | 1 |
| Functions | *yun4dong4* "to exercise" | 15 |

In terms of frequencies of each group, the metaphor POLITICS IS SPORT focuses on the mappings related to the "functions" in sport. In other words, this case does not pay attention on qualifying to the "competing" aspect or referring to "golf or tennis", "win/loses", and "rules". Instead, this metaphor emphasizes the conception of "exercising" of political power. Thus, the Mapping Principle can be as (5),

**(5) Mapping Principle for POLITICS IS SPORT**

Politics is understood as sport because sport involves physical exercising and politics involves mental exercising.

Through investigating the mapping principles of the five metaphors, we can say that this corpus-based method is better than the original one (Ahrens 2002) in generating mapping principles of metaphors because only this method is able to provide the quantitative information of the correspondence between the source and target domains as well as to determine the salient and significant mappings for each conceptual metaphor.

## 3. Conclusion

This study uses a more convincing and quantitative method to explore how to generate Mapping Principles between source and target domains. Even though Ahrens' (2002) Conceptual Mapping Model has systematic ways to generate Mapping Principles, it is generated from native speakers' intuition, which lacks empirical evidence to support whether it truthfully reflects the correspondences existing between source and target domains in metaphorical expressions. Alternatively, this corpus-based method can provide quantitative way to generate Mapping Principle existing in real usages of metaphorical expressions.

Our method is supported by five metaphor analyses: POLITICS IS BUILDING, POLITICS IS A JOURNEY, POLITICS IS A PLAY, POLITICS IS A COMPETITION and POLITICS IS SPORT. The corpora data show that the underlying reason the target domain of "politics" selects the source domain of "building" to emphasize the concept of "structure"; it selects the source domain of "journey" to emphasize the notion of "traveling through

roads/routes"; it selects the source domain to borrow conceptualization of "performance to the public"; it selects "competition" for emphasizing the notion "conflict"; it selects "sport" as a source domain to emphasize the concept of "exercising".

In the future, we will follow Ahrens et al's (2003) proposal to integrate the Mapping principles with SUMO (i.e. Suggested Upper Merged Ontology) to restrict these mapping principles. In particular, we would like to check the inference rules of the source domains "competition", "war" and "sport" or "business" and "risk" in the political metaphors and try to figure out what mapping principles can be merged or subsumed under particular superordinate domains.

**Acknowledgement**

I thank Prof. Kathleen Ahrens and two anonymous reviewers for their valuable comments. I am responsible for any errors in my paper. E-mail for correspondence: d91142001@ntu.edu.tw.

**References**

Ahrens K, Chung S-F, and Huang C-R (2003). Conceptual Metaphors: Ontology-based representation and corpora driven Mapping Principles. *Proceedings of the ACL Workshop on the Lexicon and Figurative Language*. Sapporo, Japan. 53-41.

Ahrens K. (2002). When love is not digested: underlying reasons for source to target domain pairings in the contemporary theory of metaphor. *Proceedings of the First Cognitive Linguistics Conference*. Cheng-Chi University. 273-302.

Chung, S-F, Ahrens, K. and Huang C-R (2003) ECONOMY IS A PERSON: A Chinese-English Corpora and Ontological-based Comparison Using the Conceptual Mapping Model. To appear in *the Proceedings of the 15th ROCLING Conference for the Association for Computational Linguistics and Chinese Language Processing*, National Tsing-Hwa University, Taiwan.

CKIP (1995). *Technical Report no. 95-02*. Academic Sincia Press. Taipei: Nankang.

Lakoff G. (1993). The contemporary theory of metaphor. In Andrew Ortony (eds.) *Metaphor and Thought* (2nd ed). Cambridge: Cambridge University Press. 202-251.

Lakoff, G, & Johnson, M. (1980). *Metaphors We Live by*. Chicago: Chicago University Press.

# Extracting Verb-Noun Collocations from Text

**Jia Yan Jian**
Department of Computer Science
National Tsing Hua University
101, Kuangfu Road, Hsinchu, Taiwan
g914339@oz.nthu.edu.tw

## Abstract

In this paper, we describe a new method for extracting monolingual collocations. The method is based on statistical methods extracts. VN collocations from large textual corpora. Being able to extract a large number of collocations is very critical to machine translation and many other application. The method has an element of snowballing in it. Initially, one identifies a pattern that will produce a large portion of VN collocations. We experimented with an implementation of the proposed method on a large corpus with satisfactory results. The patterns are further refined to improve on the precision ration.

## 1 Introduction

Collocations are recurrent combinations of words that co-occur more often than chance. Collocations like terminology tend to be lexicalized and have a somehow more restricted meaning than the surface form suggested (Justerson and Katz 1994). The words in a collocation may be appearing next to each other (rigid collocation) or otherwise (flexible/elastic collocations). On the other hand, collocations can be classified into lexical and grammatical collocations (Benson, Benson, Ilson, 1986). Lexical collocations are formed between content words, while the grammatical collocation has to do with a content word with a function word or a syntactic structure. Collocations are pervasive in all types of writing and can be found in phrases, chunks, proper names, idioms, and terminology.

Automatic extraction of monolingual and bilingual collocations are important for many applications, including Computer Assisted Language Learning, natural language generation, word sense disambiguation, machine translation, lexicography, and cross language information retrieval. Hank and Church (1990) pointed out the usefulness of pointwise mutual information for identifying collocations in lexicography. Justeson and Katz (1995) proposed to identify technical terminology based on preferred linguistic patterns and discourse property of repetition. Among many general methods presented in Manning and Schutze (1999), the best method is filtering based on both linguistic and statistical constraints. Smadja (1993) presented a program called XTRACT, based on mean and variance of the distance between two words that is capable of computing flexible collocations. Kupiec (1992) proposed to extract bilingual noun phrases using statitistical analysis of coocurrance of phrases. Smadja, McKeown, and Hatzivassiloglou (1996) extended the EXTRACT approach to handling of bilingual collocation based mainly on the statistical measures of Dice coefficient. Dunning (1993) pointed out the weakness of mutual information and showed that log likelihood ratios are more effective in identifying monolingual collocations especially when the occurrence count is very low.

Smadja's XTRACT is the seminal work on extracting collocation types. XTRACT invloves three different statistical measures related to how likely a pair of words is part of a collocation type. It is complicated to set different thresholds for each of these statistical measures. We decided to research and develop a new and simpler method for extracting monolingual collocations. We describe the experiments and evaluation in Section 3. The limitations and related issues will be taken up in Section 4. We conclude and give future direction in Section 5.

## 2 The algorithm

We used Sinorama Corpus to develop methods for extracting monolingual collocations. A number of necessary preprocessing steps were carried out. Those preprocessing steps include:

1. Part of speech tagging for English and Chinese test
2. N-gram construction
3. Logarithmic likelihood ratio (LLR) computation

**Log-likelihood ratio : LLR(x;y)**

$$LLR(x,y) = -2\log_2 \frac{p_1^{K_1}(1-p_1)^{n_1-k_1}(1-p_2)^{n_2-k_2}}{p^{k_1}(1-p)^{n_1-k_1}p^{k_2}(1-p)^{n_2-k_2}}$$

$k_1$ :# of pairs that contain x and y simultaneously.
$k_2$ :# of pairs that contain x but do not contain y.
$n_1$ :# of pairs that contain y
$n_2$ :# of pairs that does not contain y
$p_1 = k_1/n_1$, $p_2 = k_2/n_2$,
$p = (k_1+k_2)/(n_1+n_2)$

### 2.1 Extraction of English VN collocations

In our research, we discovered some problems about XTRACT. The problems with XTRACT include:

1. XTRACT produce a list of collocation types rather than instances.
2. XTRACT is complicated because it requires thresholds for three statistical measures.
3. There is no systematic way of setting thresholds for a certain level of confidence.
4. XTRACT is based on the author's intuition about collocation.
5. XTRACT does not provide explicitly types of collocation.

For the above reasons, we decided to research and explore new methods for extracting monolingual collocations.

#### 2.1.1 Step1: Computing such VN types with high counts

The method has an element of snowballing in it. Initially, one identifies a pattern that will produce a large portion of VN collocation. We started with the following pattern(1):

**V + ART or POSS + ⋯ + N** **(1)**

By extracting such VN types with high counts, we got a list of highly likely collocation types. In addition, we also take the passive form(2) of VN into consideration:

**ART or POSS + N + ⋯ + be + Ved (the passive VN)** **(2)**

The list is further filtered for higher precision: the pairs with LLR lower than 7.88 (confidence level 95%) are removed from consideration.

#### 2.1.2 Step2: Extracting VN patterns from corpus

After obtaining the list, we gather all the instances where the VN appears in the corpus. From the instances, we compute the following patterns(3) for extracting VN collocations:

**POS preceding V**
**POS sequence between V and O** **(3)**
**POS following O**

and we also consequently consider the passive form and its context:

**POS preceding O**
**POS sequence between O and V** **(4)**
**POS following V**

### 2.1.3 Step3: Manipulating the correct structure statistics of VN patterns

We eliminated patterns that appear less than three times. These patterns are much more stringent than pattern we started out with. These patterns help us get rid of unlikely VN instances such as "make film" in "make a leap into TV and film," since the POS sequence of "a leap into TV and" has a low count in the initial batch of "likely" collocations. On the other hand, "make film" in "make my first film" would be kept as a legistimate instance of VN, since the pos sequence of "my first" has rather high count in the initial batch of "likely" collocations.

Actually, the POS sequences of intervening words has a skew distribution concentrating on a dozen of short phrases(see Table1):

**Table 1**
Samples of VN collocation from text

| VN collocation | Translation | POS of VN |
|---|---|---|
| **ride a bike** | **騎自行車** | **vb + at + nn** |
| **take my advice** | **聽我的勸告** | **vb + pp$ + nn** |
| **keep a diary** | **寫日記** | **vb + at + nn** |
| **action will be taken** | **採取行動** | **nn + md + be + vbd** |
| **problem is solved** | **解決問題** | **nn + be + vbd** |
| **decision can be made** | **做決定** | **nn + md + be + vbd** |

These patterns can be coupled with other constraints for best results:

1. No punctuation marks should come between V and O
2. The noun closest to the verb takes precedence

For now, we only consider verbs with two obligatory arguments of subject and object. Therefore, we exclude instance like (make, choice) in "**make entertainment** at home a **choice."** We plan to extract VN in three-argument proposition separately.

The other issue has to do with data sparseness. For collocation types with low count, the estimation of LLR is not as reliable. In the future, we will also experiment with using search engine such as Google to estimate word counts and VN instance count for more reliable estimation of LLR.

XTRACT does not touch on the issue of identify VN collocation instances in (6) and exclude that in (5). In our research, we explored the identification of collocation instances and attempt to avoid cases that maybe a correct collocation type but not a correct collocation instance.

… *make* a leap into TV and *film*… (5)
… *made* great efforts to promote documentary *film*… (6)

## 2.2 Example

To extract VN collocations, we first run part of speech tagging on sentences. For instance, we get the results of tagging below :

He/pps defines/vbz success/nn for/in a/at paper/nn as/cs not/* needing/vbg to/to exert/vb political/jj influence/nn or/cc obtain/vb financial/jj subsidies/nns ,/, but/cc rather/rb being/beg able/jj to/to rely/vb wholly/rb on/in content/nn to/to attract/vb readers/nns that/cs in/in turn/nn attract/vb advertisers/nns ,/, and/cc thus/rb keep/vb afloat/rb by/in its/pp$ own/jj efforts/nns ./.

After tagging English sentences, we construct N-gram extracted likely VN types with high count from bigram, trigram and fourgram. We then obtained got a list of highly likely collocation types (Table 2). The pairs with LLR lower then 7.88 are eliminated from Table 2. If the pair appeared less than once. we also eliminated the pair.

After obtaining likely collocation types, we gathered all instances where the VN appears in the corpus. The distance between the verb and the object is at most five words. Both of the words before the verb and after the object are recorded. Table 3 shows those patterns of VN instances.

**Table 2**
A list of highly likely collocation types

| Verb | Noun | Count (VN) | Count(V) | Count(N) | llr_score |
|---|---|---|---|---|---|
| have | influence | 24 | 5293 | 57 | 52.28961 |
| exert | influence | 4 | 14 | 57 | 40.58210 |
| exercise | influence | 4 | 23 | 57 | 36.09338 |
| reduce | influence | 3 | 188 | 57 | 12.43681 |
| eradicate | influence | 1 | 6 | 57 | 8.876641 |
| root | influence | 1 | 6 | 57 | 8.876641 |

**Table 3**
Extracting VN collocation from corpus

| Rec | V-1 | Verb | N-5 | N-4 | N-3 | N-2 | N-1 | Noun | N+1 |
|---|---|---|---|---|---|---|---|---|---|
| 96335 | 't | have | | | | | much | influence | on |
| 55203 | woman | have | | | | | some | influence | , |
| 129530 | tank | have | | | | a | considerable | influence | on |
| 122706 | He | have | | | | | an | influence | on |
| 123975 | mother | have | | | | | considerable | influence | . |
| 125192 | Wen | have | | | | a | great | influence | on |
| 9326 | which | have | | | such | a | powerful | influence | on |
| 56033 | as | have | | | | an | enormous | influence | throughout |
| 67666 | have | have | | | | | less | influence | than |
| 76130 | have | have | | | | | lasting | influence | on |
| 95098 | always | have | | | | a | certain | influence | on |
| 125182 | Xi | have | | | | the | greatest | influence | on |
| 5704 | have | have | | | a | very | negative | influence | . |
| 1742 | have | have | | a | deep | and | lasting | influence | . |
| 111368 | owner | have | | | | no | less | influence | than |
| 96654 | thus | have | | | | a | decisive | influence | on |
| 109816 | family | have | | | | the | greatest | influence | on |
| 115428 | png | have | | | be | under | foreign | influence | , |
| 39165 | to | exert | | | | | | influence | . |
| 112540 | to | exert | | | | | political | influence | or |
| 118754 | to | exert | | | | | his | influence | to |
| 106807 | to | exert | | | | a | positive | influence | for |
| 106846 | it | exert | | | a | powerful | cultural | influence | throughout |
| 46061 | whohas | exert | | | | | enormous | influence | upon |
| 123962 | best | exercise | | | | a | restrain | influence | on |
| 40774 | and | exercise | | | | her | political | influence | in |
| 127061 | to | reduce | | | | | the | influence | of |

## 3 Experiment and evaluation

We worked with around 50,000 aligned sentences from the Sinorama parallel Corpus in our experiments with an implementation of the proposed method. The average English sentence had 43.95 words. From the experimental data, we have extracted 17,298 VN collocation types. Then, we could obtain 45,080 VN instances for these VN types. See Table 3 for some examples for the verb "influence."

We select 100 sentences from the parallel corpus of Sinorama magazine to evaluate the performance. A human judge majoring in English identified the VN collocations in these sentences. The manual VN collocations are compared with the instances extracted from the corpus and the result is showed in the Appendix. The evaluation indicates an average recall rate of 74.47% and precision of 66.67 %.

**Table 4**
Experiment result of VN collocation extracted from Sinorama parallel Corpus

| #answer keys | #output | #Correct | Recall (%) | Precision (%) |
|---|---|---|---|---|
| **94** | **105** | **70** | **74.47** | **66.67** |

It is very difficult to evaluation the experimental results. There were obvious and clear-cut collocations and non collocation, but there were a lot of cases such as "improve environment" and "share housework" that were difficult to judge and may be evaluated differently by different people. There is room for improvement as far as recall and precision ratios are concerned. Nevertheless, the extracted VNs are very diverse and useful for language learning purpose.

## 4 Discussion

The proposed approach offers a simple algorithm for automatic acquisition of the VN instances from a corpus. The method is particularly interested in following ways:

i. We use a data-driven approach to extract monolingual collocations.
ii. The algorithm is applicable to elastic collocations.
iii. Systematic way of setting thresholds for a certain level of confidence
iv. We could obtained instances of VN collocation through the simple statistical information.

While Xtract extracts VN types, we focus on the VN instances. It is understandable that we would get slightly lower recall and precision rates.

## 5 Conclusion & Future work

In this paper, we describe an algorithm that employs statistical analyses to extract instance of VN collocations from a corpus. The algorithm is applicable to elastic collocations. The main difference between our algorithm and Xtract lies in that we extract the instances from the sentence instead of extracting the VN types directly.

Moreover, in our research we observe other types related to VN such as VP (ie. verb + preposition) and VNP (ie. verb + noun + preposition). In the future, we will further take these two patterns into consideration to extract more types of verb-related collocations.

## References

Benson, Morton., Evelyn Benson, and Robert Ilson. *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. John Benjamins, Amsterdam, Netherlands, 1986.

Choueka, Y. (1988) : "Looking for needles in a haystack", Actes RIAO, Conference on User-Oriented Context Based Text and Image Handling, Cambridge, p. 609-623.

Choueka, Y.; Klein, and Neuwitz, E.. Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *Journal of the Association for Literary and Linguistic Computing*, 4(1):34-8, (1983)

Church, K. W. and Hanks, P. Word association norms, mutual information, and lexicography. Computational Linguistics, 1990, 16(1), pp. 22-29.

Dagan, I. and K. Church. Termight: Identifying and translation technical terminology. In *Proc. of the 4th Conference on Applied Natural Language Processing (ANLP)*, pages 34-40, Stuttgart, Germany, 1994.

Dunning, T (1993) Accurate methods for the statistics of surprise and coincidence, Computational Linguistics 19:1, 61-75.

Haruno, M., S. Ikehara, and T. Yamazaki. Learning bilingual collocations by word-level sorting. In *Proc. of the 16th International Conference on Computational Linguistics (COLING '96)*, Copenhagen, Denmark, 1996.

Huang, C.-R., K.-J. Chen, Y.-Y. Yang, Character-based Collocation for Mandarin Chinese, In ACL 2000, 540-543.

Inkpen, Diana Zaiu and Hirst, Graeme. ``Acquiring collocations for lexical choice between near-synonyms." SIGLEX Workshop on Unsupervised Lexical Acquisition, 40th meeting of the Association for Computational Lin

Justeson, J.S. and Slava M. Katz (1995). Technical Terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9-27.

Kupiec, Julian. An algorithm for finding noun phrase correspondences in bilingual corpora. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics, Columbus, Ohio, 1993.

Lin, D. Using collocation statistics in information extraction. In *Proc. of the Seventh Message Understanding Conference (MUC-7)*, 1998.

Melamed, I. Dan. "A Word-to-Word Model of Translational Equivalence". In Procs. of the ACL97. pp 490-497. Madrid Spain, 1997.

Smadja, F. 1993. Retrieving collocations from text: Xtract. Computational Linguistics, 19(1):143-177

Smadja, F., K.R. McKeown, and V. Hatzivassiloglou. Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):1-38, 1996.

## Appendix

The manual VN collocations are compared with the instances extracted from the corpus:

| Rec | Manual VN collocations | Automatic extracting VN collocations |
|---|---|---|
| 162 | *ask question* \ hold conference \ *grant amnesty* \ *realize probability* | *grant amnesty* \ *ask question* \ *realize probability* |
| 1647 | enforce rule (被動) \ *break rule* \ *enhance image* \ forge reputation \ *respect law* | *enhance image* \ *respect law* \ improve organization \ *break rule* \ reward reputation |
| 2106 | | |
| 4898 | | |
| 5857 | take power \ *do reserch* | *done research* \ accuse linguistics |
| 6489 | make demand \ make improvement \ *make breakthrough* | *make breakthrough* |
| 6871 | *put mark* \ *release album* | *release album* \ *put mark* |
| 6887 | | meet friend |
| 7420 | *take risk* \ *make start* | *take risk* \ *make start* \ lead risk |

| | | |
|---|---|---|
| 7710 | | |
| 7878 | *make money* \ *make profit* \ *rise price* | stop conglomerate \ *make money* \ *rise price* \ *make profit* |
| 7932 | *eliminate unfariness* \ *seek equity* | *seek equity* \ *eliminate unfairness* |
| 8056 | | |
| 8510 | *improve environment* | *improve environment* |
| 8630 | | |
| 9326 | do research \ *have influcence* | *have influence* |
| 9433 | | |
| 10600 | | |
| 10624 | | contemplate footstep |
| 11293 | *understand meaning* | *understand meaning* |
| 11603 | | |
| 12937 | *receive attention* \ *witness progress* | *receive attention* \ *witness progress* |
| 13033 | *promote idea* \ *invest effort* \ *share housework* \ *expend effort* | *expend effort* \ *share housework* \ *promote idea* \ *invest effort* |
| 13491 | | |
| 13576 | | test wisdom |
| 15349 | take paycut \ exceed budget \ *unload property* | show increase \ house price \ *unload property* |
| 16949 | | |
| 17106 | block view \ *make offering* | *make offering* |
| 17608 | *lose ability* | *lose ability* \ save forest |
| 17924 | take effort \ take time | consider success |
| 18183 | | |
| 18717 | *carry work* | *carry work* |
| 18745 | | |
| 19735 | *bear son* | *bear son* |
| 20002 | *make money* \ *think way* | *make money* \ *think way* |
| 21450 | | buy portion |
| 21663 | live life | live space |
| 22610 | | |
| 23067 | *adopt method* | *adopt method* |
| 23074 | | |
| 24307 | *move production* | *move production* \ develop computer |
| 25478 | | |
| 26030 | *make thing* | *make thing* |
| 28303 | *increase chance* \ *increase production* | *increase chance* \ *increase production* |
| 28336 | | |
| 28417 | *write essay* | *write essay* |
| 28806 | *write seller* | *write seller* |
| 28826 | | |
| 29003 | *make money* \ *take care* \ *have time* | *take care* \ *make money* \ *have time* |
| 29292 | | |
| 29736 | *damage environment* | *damage environment* \ insure recovery \ choose styrofoam \ recover styrofoam |
| 30881 | *donate kidney* \ *implant kidney* | *donate kidney* \ *implant kidney* |
| 31096 | *drive car* \ take transportation \ *have responsibility* | *drive car* \ consume pastry \ *have responsibility* \ wrap candy |
| 32975 | *instruct student* | *instruct student* |

| | | |
|---|---|---|
| 33558 | *take part in* | *take part* \ detail research |
| 33993 | | |
| 33994 | *have chance* | *have chance* |
| 34008 | | excite pupil |
| 34966 | *have drink* \ *kick habit* | carry card \ ask carrier \ *have drink* \ *kick habit* |
| 35113 | | come face |
| 35898 | *announce approval (被動)* \ *bear child* | *announce approval* \ *bear child* |
| 35906 | *make adjustment* \ build contact | *make adjustment* |
| 36931 | *apply concept* | *apply concept* |
| 36988 | | supplant worth |
| 37025 | start movement | |
| 37811 | *hear sound* | *hear sound* |
| 37835 | *dedicate life* \ achieve dream (被動) \ *put effort* | *put effort* \ *dedicate life* |
| 37916 | gain influence \ *spend day* | *spend day* |
| 38197 | unload burden \ pursue success | |
| 38200 | | |
| 38231 | | begrudge money |
| 38626 | | |
| 40823 | do service | |
| 40873 | *pay attention* \ *put emphasis* \ incite response | *pay attention* \ *put emphasis* |
| 41102 | | |
| 41383 | | exist nativism |
| 41532 | | move oxcart |
| 43027 | | personalize book |
| 43199 | *follow road* | *follow road* |
| 43304 | *derive satisfaction* | *derive satisfaction* |
| 43465 | | |
| 44052 | | |
| 44189 | | strip circle |
| 44276 | *impose sanction* | *impose sanction* \ endanger specie |
| 44351 | *carry burden* \ raise image | *carry burden* |
| 44990 | | |
| 45187 | | |
| 45191 | | |
| 45499 | *pay a visit to* | *pay visit* |
| 45756 | | stoop frame |
| 45857 | *point way* | *point way* |
| 45905 | | |
| 46466 | | |
| 47134 | *offend policeman* | borrow hairpin \ *offend policeman* |
| 47226 | | |
| 47337 | | |
| 47428 | receive treatment | |
| 47720 | | |
| 48694 | | |
| 48919 | | elapse step |

# Bilingual Sentence Alignment Based on Punctuation Marks

**Kevin C. Yeh**
**g904307@oz.nthu.edu.tw**
**Department of Computer Science**
**National Tsing Hua University**

**Abstract**

We present a new approach to aligning English and Chinese sentences in parallel corpora based solely on punctuations. Although the length based approach produces high accuracy rates of sentence alignment for clean parallel corpora written in two Western languages such as French-English and German-English, it does not fair as well for parallel corpora that are noisy or written in two distant languages such as Chinese-English. It is possible to use cognates on top of length-based approach to increase alignment accuracy. However, cognates do not exist between two distant languages, therefore limiting the applicability of cognate-based approach. In this paper, we examine the feasibility of using punctuations for high accuracy sentence alignment. We have experimented with an implementation of the proposed method on the parallel corpus of Chinese-English Sinorama Magazine Corpus with satisfactory results. We also demonstrated that the method was applicable to other language pairs such as English-Japanese with minimal additional effort.

## 1. Introduction

Recently, there are renewed interests in using bilingual corpus for building systems for statistical machine translation (Brown et al. 1988, 1991), including data-driven machine translation (Dolan, Pinkham, and Richardson 2002), computer-assisted revision of translation (Jutras 2000) and cross-language information retrieval (Kwok 2001). It is therefore useful for the bilingual corpus to be aligned at the sentence level with very high precision (Moore 2002; Chuang, You and Chang 2002, Kueng and Su 2002). After that, further analyses such as phrase and word alignment, bilingual terminology extraction can be performed (Melamed 1997).

Much work has been reported in the literature of computational linguistics studying how to align English-French and English-Germany sentences. While the length-based approach (Church and Gale 1991; Brown et al. 1992) to sentence alignment produces surprisingly good results for the language pair of French and English at success rates well over 96% by sentence, it does not fair as well for alignment of English and Chinese sentences. Work on sentence alignment of English and Chinese texts (Wu 1994), indicates that the lengths of English and Chinese texts are not as highly correlated as in French-English task, leading to lower success rate (85-94%) for length based aligners. Simard, Foster, and Isabelle (1992) proposed

using cognates on top of length-based approach to improve on accuracy. They use an operational definition of cognates, which include digits, alphanumerical symbols, punctuations and alphabetical words.

Simard, Foster, and Isabelle (1992) pointed out cognates in two close languages such as English and French can be used to measure the likelihood of mutual translation. Those cognates include alphabetic words, numeric expressions, and punctuations that are almost identical and readily recognizable by the computer. However, for distant language pairs such as Chinese and English, there are no orthographic, phonetic or semantic cognates in existence, which are readily recognizable by the computer. Therefore, the inexpensive cognate-based approach is not applicable to the Chinese-English tasks. Since both of the length and cognate-based methods do not present satisfactory alignment results for distant bilingual pairs, we are motivated to find other alternative evidence that two blocks of texts are mutual translation. It turns out that punctuations can be telling evidences, if we do more than hard matching of punctuation and take into consideration of intrinsic sequencing of punctuation in ordered comparison.

## 2. Punctuation and Sentence Alignment

We will show that punctuations in Chinese and English mark texts with similar semantic properties, therefore, it is very effective to use them to measure the likelihood of mutual translation for a pair of texts.

**Punctuation Translation probability** **Punctuation Fertility probability**

| English Punctuation | Chinese Punctuation | Number of counts | Probability |
|---|---|---|---|
| ) | ) | 4 | 0.9972 |
| ," | 」, | 6 | 0.9564 |
| ." | 」, | 3 | 0.9164 |
| ! | ！ | 38 | 0.8835 |
| , | ， | 541 | 0.8099 |

| Punctuation Type | Number of Counts | Probability |
|---|---|---|
| 0-1 | 588.0005 | 0.225027 |
| 1-0 | 286.001 | 0.109452 |
| 1-1 | 1698.076 | 0.649852 |
| 1-2 | 2.466198 | 0.000944 |
| 2-1 | 0.965034 | 0.000369 |
| 2-2 | 37.19216 | 0.014233 |
| 3-1 | 0.314022 | 0.000121 |

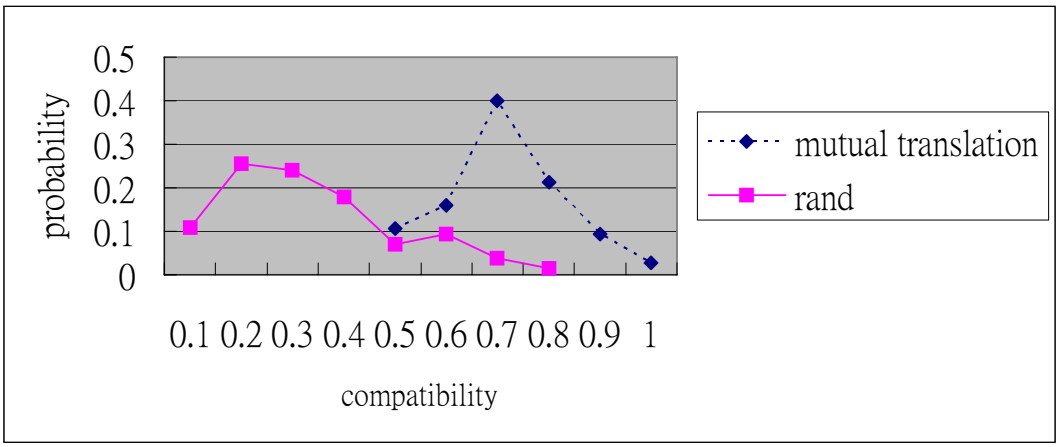
In order to explore the relationship between the punctuations in pairs of Chinese and English sentences that are mutual translations, we selected a small set of manually aligned texts and investigated the characteristics and the associated statistics between the punctuations. Following next a list of the number of counts and the probability relating the Chinese punctuation and the English punctuation was tallied. The information was then used to bootstrap on a larger corpus where an unsupervised EM algorithm and Dynamic programming are used to optimize the punctuation correspondence between a text and its translation counterpart. The EM algorithm converges quickly after the second round of training. Some of the results of the EM training are shown in the above tables.

The probability of the one-to-one match type is about 0.65, which implies that there is a large discrepancy of the punctuation mappings between Chinese and English. The punctuation compatibility is measured by using a relatively larger corpus – around one thousand articles from the Sinorama magazine. Simard, Foster and Isabelle proposed the cognate based approach as a new way of measuring how two pieces of text are mutual translation. The punctuation compatibility as an indicator of mutual translation is defined as

$$\gamma = \frac{c}{\max(n, m)}$$

where $\gamma$ = punctuation compatibility,
$c$ = number of cognate,
$n$ = number of Chinese punctuations marks
$m$ = number of English punctuation marks

We then took the aligned English-Chinese sentences that have same number of punctuation count (that is the denominator of the equation), take ten counts for example, in order to evaluate how well punctuations work as indicator of mutual translation of English and Chinese sentences. We also took the same English sentences and matched them up with randomly selected Chinese sentences to calculate the compatibility of punctuation marks of unrelated texts.




Results indicated that the average compatibility for pairs of sentences, which are mutual translations, is about 0.67 (with a standard deviation 0.170), while the average compatibility of random pairs of bilingual sentences is 0.34 (with a standard deviation 0.167). The above figure shows compatibility based on punctuation count of ten. The results indicate as the number of punctuations increases the reliability of the compatibility function is more informative. Overall, if the punctuations are soft matched in ordered comparison across the two languages, they indeed provide useful information for effective sentence alignment.

We define a probability of the sequence of punctuations $E_i$ in one language (L1) translating to the sequence $C_j$ of punctuation in another language (L2) as follows:

$$P(E_i, C_j) = \prod_{k=1,m} P(p_k, \pi_k) P(|p_k|, |\pi_k|)$$

where $p_k$ and $\pi_k$ is one or two punctuations,
$p_1 p_2 \cdots p_m = E_i$, the English punctuations,
$c_1 c_2 \cdots c_m = C_j$, the Chinese punctuations,
$|p_k|$ and $|\pi_k|$ are the number of punctuations in $p_k$ and $\pi_k$ respectively,
P($p_k$, $\pi_k$) = probability of $p_k$ translating into $\pi_k$,
P($j$, $k$) = probability of $j$ punctuations in L1 translating into punctuation in L2.

We observed that in most cases the links of punctuations do not cross each other much like the situation with sentence alignment. Therefore, it is possible to use the dynamic programming procedure to soft match the punctuations across languages, finding the Viterbi path as long as we have the punctuation translation function P($p_k$, $\pi_k$) and the fertility function P(j, k).

Not like the way Simard et al. (1992) handled cognates, we model the compatibility of punctuations across two languages using Binomial distribution. We model the problem as each punctuation appearing in one language either has a

counterpart across translation or not. And for each punctuation, the probability of having a translation counterpart is independent with a fixed value of p.

We differ from Simard approach in the following interesting ways. First, we use the accumulative value of Binomial distribution, while Simard et al. used a likelihood ratio.Second, we go beyond mere hard matching and allow a punctuation mark in one language to match up with a number of compatible punctuations. The compatibility is modeled based on the lexical translation probability proposed by Brown et al. (1991).Finally, we take into consideration of intrinsic sequencing of punctuation in ordered comparison, the flexible and ordered comparison of punctuation is carried out via dynamic programming.

Following Gale and Church (1991), we appeal to Bayes Theorem to estimate the likelihood of aligning two text blocks E and C by calculating P(E, C) P(match).We adopt the same dynamic programming method, but use punctuations to measure the likelihood of mutual translation instead of lengths. For that we define the probability P(E, C) that two text blocks E and C are mutual translation as follows: Given two blocks of text E and C, we first strip off non-punctuations therein to get the punctuations strings $E_i$ and $C_j$ and find out the maximum number of punctuations n.

Subsequently, the dynamic programming procedure mentioned before is carried out to find out the value of r, the number of compatible punctuations in ordered comparison of punctuations across languages. Therefore we have:

$$P(E,C) = \sum_{k=1}^{t} P(m_k) P(E_{i,k}, C_{j,k}) = \sum_{k=1}^{t} \mathrm{P(m_k)} b(r_k, n_k)$$

$$= \sum_{k=1}^{t} \mathrm{P(m_k)} \left( \frac{n_k}{r_k} \right) {p_k}^{n^k} (1 - p_K)^{r^k - n^k}$$

where $n_K$ = the number of compatible punctuations in ordered comparison,

$r_k$ = the max number of punctuations from English text or Chinese text

$p_K$ = the probability of existence of a compatible punctuation across from one language to the other.

$P(m_K)$ = the match type probability aligning $E_{i,k}$ and $C_{j,k}$

From the data, we have found that about two third of the times, a sentence in one language matches exactly one sentence in the other language (1-1). Other additional possibilities are also considered: 1-0 (including 0-1), and many-1 (including 1-many). Chinese-English parallel corpora are considerably noisier, reflecting from wider

possibilities of match types. Here we used the same probabilistic figures as proposed in Chuang and Chang (2002). The following table shows all eight possibilities used in our implementation.

| Match type | 1-1 | 1-0, 0-1 | 1-2 | 2-1 | 1-3 | 1-4 | 1-5 |
|---|---|---|---|---|---|---|---|
| Probability | 0.65 | 0.000197 | 0.0526 | 0.178 | 0.066 | 0.0013 | 0.00013 |

### 3.1. First Experiment and Evaluation

In the first experiment, we assessed the performance of punctuation-based sentence alignment, we have randomly selected five bilingual articles from three different bilingual corpora to test out to an implementation of the proposed method. Evaluation of the experiment results were made by native Chinese college students with good knowledge in English. Some experimental results of sentence alignment based on length and punctuation are shown in Appendix (Table A). Shaded parts indicate imprecision in alignment results. We calculated the precision rates by dividing the number of un-shaded sentences (counting both English and Chinese sentences) by total number of sentences proposed. Since we did not exclude aligned pair using a threshold, the recall rate should be the same as the precision rate. The experimental results indicate that when non 1-1 matches next to each other tend to fail the length-based aligner. However, the punctuation-based aligner appears to handle such cases more successfully.

**Precision Evaluation using punctuations**

| Articles | Length-only | Punctuation-only | Improvement |
|---|---|---|---|
| World in a box* | 91.5 | 98.8 | 7.3 |
| What clones* | 86.5 | 96.6 | 10.1 |
| New University** | 93.0 | 95.3 | 2.3 |
| Book I-2 *** | 96.5 | 98.9 | 2.4 |
| Book II-8 *** | 97.1 | 98.0 | 0.9 |

* Scientific American Magazine

** Sinorama Magazine

*** Harry Porter

### 3.2. Second Experiment and Evaluation

In the second experiment, we evaluated our method testing on a larger range of corpus data. We used all the English and Chinese articles of Scientific American Corpus from January 2003 to December 2003. There are 67 articles, 1523 English sentences, and 1599 Chinese sentences. All the articles include both the English text and their corresponding Chinese counterpart. Here are the results of the experiment:

| | Precision | Recall |
|---|---|---|

| | | |
|---|---|---|
| Excluding partially incorrect and missing errors | 95.8% | 100.0% |
| Including partially incorrect and missing errors | 93.0% | 98.2% |

## 4. Conclusion

We developed a very effective sentence alignment method based on punctuations. The probability of the match between different punctuation marks between the source and the target is calculated based on large bilingual corpora. The punctuation alignment has the property of a binomial distribution. We have experimented with an implementation of the proposed method on a large parallel corpus data. The experiment results show that the punctuation- based approach outperforms the length-based approach with precision rates approaching 93%.

We have explored ways of extending punctuation-based method. First, there is possibility that we may want to interleave the matching of punctuations and regular text segments between punctuations for sub-sentential alignment. We observed that although word alignment links do cross one and other a lot, they general seem not to cross the links between punctuations. Also since the method is quite general, it would be interesting to see if one can adapt the method to handle other language pairs. We have hand-coded a small English-Japanese punctuation mapping table, and convert our alignment program to handle Alignment of Japanese and English texts. It appears that the adapted program works with compatible performance to the original one. Please see example in Appendix (Table B).

A number of interesting future directions present themselves. First, punctuation alignment can be exploited to constrain word alignment and reduce error rates. Second, the punctuation alignment make possible a finer-grained level of bilingual analysis and can provide a strikingly different translation memory and bilingual concordance for more effective example-based machine translation (EBMT), computer assisted translation and language learning (CAT and CALL).

## References

Brown, P. F., J. C. Lai and R. L. Mercer (1991), 'Aligning sentences in parallel corpora', in 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, USA. pp. 169-176.

Chuang, T., G.N. You, J.S. Chang (2002) Adaptive Bilingual Sentence Alignment, Lecture Notes in Artificial Intelligence 2499, 21-30.

Dolan, William B., Jessie Pinkham & Stephen D. Richardson (2002) MSR-MT: The Microsoft Research Machine Translation System, AMTA 2002, 237-239.

Gale, William A. & Kenneth W. Church (1993), A program for aligning sentences in bilingual corpus. In Computational Linguistics, vol. 19, pp. 75-102.

Jutras, J-M 2000. An Automatic Reviser: The TransCheck System, In Proc. of Applied Natural Language Processing, 127-134.

Kueng, T.L. and Keh-Yih Su, 2002. A Robust Cross-Domain Bilingual Sentence Alignment Model, In Proceedings of the 19th International Conference on Computational Linguistics.

Kwok, KL. 2001. NTCIR-2 Chinese, Cross-Language Retrieval Experiments Using PIRCS. In Proceedings of the Second NTCIR Workshop Meeting, pp. (5) 14-20, National Institute of Informatics, Japan.

Melamed, I. Dan (1997), A portable algorithm for mapping bitext correspondence. In The 35th Conference of the Association for Computational Linguistics (ACL 1997), Madrid, Spain.

Moore, Robert C., 2002. Fast and Accurate Sentence Alignment of Bilingual Corpora., AMTA 2002, 135-144.

Simard, M., G. Foster & P. Isabelle (1992), Using cognates to align sentences in bilingual corpora. In Proceedings of TMI92, Montreal, Canada, pp. 67-81.

Wu, Dekai (1994), Aligning a parallel English-Chinese corpus statistically with lexical criteria. In The Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, New Mexico, USA, pp. 80-87.6. Tainan: National Cheng Kung University.

# Appendix

**Table A. Experimental result of sentence alignment based on length and punctuation.**

| Sentence alignment based on length | | |
|---|---|---|
| Type | English text | Chinese Text |
| 12 | Allowing education to be led by the market may also lead to deficiencies in teaching practices. | 市場領導教育還可能引發教學上的弊病。台大法律系教授賀德芬說，對法律系學生來說，考上司法官、高考是最好的出路， |
| 11 | Professor He Te-fen of NTU's Department of Law say that for law students, the best opportunity for advancement is to pass the recruitment examinations for public prosecutors and judges, or the senior civil service exams. | 「有些學生上課只想具體知道如何答考題，選課標準就是老師的教書方式是不是對考試有用。」 |
| 31 | "In class, some students only want to learn specifically how to answer exam questions, and their choice of courses depends on whether the instructor's teaching method is helpful for passing the exams." Some instructors, seeing that some students do not take good notes, even designate one who does to give them to the others for reference. But this results in most of the students taking no notes at all, because after all they will get photocopies, paid for out of the class expenses fund. | 甚至有老師因爲看學生的筆記記不好，指定做得好的同學給其他人參考，以提高系上的錄取率，結果變成學生也不做筆記了，反正有班費可以影印給大家，這個現象還有個名詞叫「共筆」。 |
| Sentence alignment based on punctuation | | |
| 11 | Allowing education to be led by the market may also lead to deficiencies in teaching practices. | 市場領導教育還可能引發教學上的弊病。 |
| 11 | Professor He Te-fen of NTU's Department of Law say that for law students, the best opportunity for advancement is to pass the recruitment examinations for public prosecutors and judges, or the senior civil service exams. | 台大法律系教授賀德芬說，對法律系學生來說，考上司法官、高考是最好的出路， |
| 11 | "In class, some students only want to learn specifically how to answer exam questions, and their choice of courses depends on whether the instructor's teaching method is helpful for passing the exams." | 「有些學生上課只想具體知道如何答考題，選課標準就是老師的教書方式是不是對考試有用。」 |
| 21 | Some instructors, seeing that some students do not take good notes, even designate one who does to give them to the others for reference. But this results in most of the students taking no notes at all, because after all they will get photocopies, paid for out of the | 甚至有老師因爲看學生的筆記記不好，指定做得好的同學給其他人參考，以提高系上的錄取率，結果變成學生也不做筆記了，反正有班費可以影印給大家，這個現象還有個名詞叫「共筆」。 |

| | | |
|---|---|---|
| | class expenses fund. | |

**Table B. English-Japanese sentence alignment example.**

| Sentence alignment based on punctuation | | |
|---|---|---|
| Type | English text | Japanese Text |
| 12 | It turns out that about two-thirds of the names examined were suitable for either women or men.<br>. | その結果、３分の２の名前が男でも女でも通用するものであることがわかった。漢代の女性の名前には実に力強いものも少なくない。<br>、。。 |
| 21 | Wang Mang, who usurped the throne in 9 AD, named his daughter Jie ("nimble and quick"). The daughter of the emperor Huan Di (132-167 AD) was named Jian ("solid and resolute") while her mother, the empress Deng, had the even more emphatic name of Mengnu, which means "fierce woman"!<br>, , ( " " ) . ( ) ( " " ) , , , " " ! | 王莽の娘の名は「倢」、後漢の桓帝の娘の名は「堅」といい、桓帝の時の皇后の名は、より直接的な「猛女」というものだったのである。<br>「」、「」、、「」。 |
| 11 | Says Liu, "These names show that society at that time had not yet come to hold the two sexes to such very different standards."<br>, " . " | 「この現象は、男性と女性の道徳行爲に対する社会の要求が、あまり違わなかったことを示しています」と劉増貴さんは言う。<br>「、、」。 |