Chinese Sketch Engine and Mapping Principles: A Corpus-Based Study of Conceptual Metaphors Using the BUILDING Source Domain

Shu-Ping Gong
National Taiwan University
1 Roosevelt Road, Section 4
Taipei 106 Taiwan
d91142001@ntu.edu.tw

Kathleen Ahrens
National Taiwan University
1 Roosevelt Road, Section 4
Taipei 106 Taiwan
kathleenahrens@yahoo.com.

Chu-Ren Huang
Academia Sinica
128, Sec. 2, Academia Road
Nankang, Taipei 115 Taiwan
churen@sinica.edu.tw

Abstract

The goal of this paper is to use a large-scale corpus, i.e. the Gigaword Corpus via the interface of Chinese Sketch Engine, to determine underlying reasons between source and target domain pairings for conceptual metaphors, called Mapping Principles. In particular, we will employ a frequency-based collocational approach to examine metaphors that use the source domain of BUILDING in Mandarin Chinese. The corpus data demonstrates that different target domains use the source domain of BUILDING for different underlying reasons. Our study follows the contrastive linguistic analysis of conceptual metaphors proposed by the Conceptual Mapping Model and help us better understand why certain mappings exist between knowledge domains.

1 Introduction

Ahrens (2002) proposed an intuition-based approach to analyze metaphoric expressions in terms of the entities, qualities and functions that can map between a source and a target domain. Her study relied on native speakers' intuition to generate metaphoric expressions and determine the underlying reason that is mapped conceptually from the source to the target domain, called Mapping Principle. For example, in the following example from the metaphor IDEAS ARE BUILDINGS, i.e. *nide lundian genji shi sheme* "What is the foundation of your argument?", the Mapping Principle (MP) is postulated: *An idea is understood as a building because a building involves a physical structure and ideas involve abstract structure*.

The underlying reasons (or Mapping Principles) for the source-target domain pairings from linguistic data allow predictions to be made concerning processing conventional and novel metaphors (Ahrens, 2002). Her off-line psycholinguistic studies demonstrated that novel metaphors that follow MPs were rated more acceptable or interpretable than conventional metaphors. Additionally, novel metaphors that do not follow MPs were rated less acceptable and interpretable than novel metaphors that do follow MPs.

In addition, in order to verify the mapping principles, a corpora-based method has been developed (Ahrens, Chung & Huang 2003, 2004; Huang, Chung, & Ahrens 2006). In particular, Ahrens, Chung & Huang (2003, 2004) integrated the Conceptual Mapping Model (Ahrens 2002) with an ontology-based knowledge representation (i.e. SUMO) and WordNet to verify mapping principles between target-source domain pairings. They proposed that each source-target domain pairing has a prototypical instance of mapping as indicated by a lexical item that is frequently mapped, as compared with other mappings. Ahrens et al. (2003) defined two numerical criteria for the MP determination. The first criterion is ten metaphoric instances as the minimal number of tokens for the MP determination. The second numerical criterion is that thirty percentages of tokens is required to reach consensus for a mapping principle. For example, 38% (i.e. 39 tokens) of 102 instances are mapped with the lexical item *jianshe* "construction" for the metaphor ECONOMY IS A BUILDING and verify the following Mapping Principle: *Economy is a building because buildings involve a structure and economy involves an (abstract)*

structure. Another solution, i.e. the lexicon-defined method via the WordNet and the SUMO, is proposed (Ahrens et al. 2004) when conventional metaphor examples are fewer than 10 tokens, or when the percentage of a single lexical item is less than 30%. For example, they found that five metaphoric instances for LOVE IS PLANT are defined as growth three times out of five examples in the WordNet senses and the SUMO category. The information from the WordNet and SUMO can then verify the Mapping Principle for LOVE IS PLANT: *Love is understood as plant because plants involve physical growth and love involves emotional growth*.

However, most current lexical resources are too small to acquire enough conventional metaphor examples. A small-scale corpus, such as the Academia Sinica Corpus, will not meet researchers' need in studying metaphors. Linguists can't draw out any theories based on a few metaphoric examples because a small number of tokens can not show any systematic linguistic patterns. In Ahrens et al.'s (2004) study, for example, there are only two conceptual metaphors out of twelve conceptual metaphors that have more than 100 metaphoric tokens, i.e. ECONOMY IS A PERSON and ECONOMY IS A BUILDING. However, there are only forty metaphoric tokens for ECONOMY IS COMPETITION and twenty-three metaphoric instances for ECONOMY IS WAR, etc. We are afraid that the small corpus data of metaphor examples can't reflect the authentic patterns when people use metaphors in discourse. So, the linguistic patterns will be more convincing when metaphor models and theories are postulated based on more than one hundred metaphoric examples. Therefore, we need a large-scale corpus to find enough conventional metaphor examples in order to establish the linguistic patterns or rules by means of the occurrence frequency.

In addition, the traditional method (Ahrens et al. 2003, 2004) is to determine the underlying reasons for a target domain to use a source domain (Ahrens et al. 2003, 2004; Ahrens 2002). However, this method limits the opportunity to examine the different underlying reasons for different target domains to select the same source domain. It is possible that different target domains use a source domain to highlight different aspects of the source domain. For example, Ahrens (2002) found that both lexical items "ideas" and "love" are described in terms of the source domain of FOOD. However, the two concepts select distinct aspects of FOOD for the Mapping Principles. The aspect of digestion is emphasized for IDEA IS FOOD while the dimension of taste is highlighted for LOVE IS FOOD. This comparison example suggests that even though different target domains repeatedly use the same source domain, they may select different aspects of the source domain for distinct underlying reasons. Therefore, it is necessary to use the method from the opposite direction, i.e. the way to examine metaphors from the source to the target domain, which may allow one to expand the numbers of target domains that use the same source domain and to better understand how the source domain contributes to metaphoric meanings when the different target domains select the same source domain.

In this study, we are going to use a large-scale corpus, i.e. the Gigaword Corpus via the interface of the Chinese Sketch Engine, to determine the mapping principles between source and target domain pairings in Mandarin Chinese. In particular, we employ a frequency-based collocational approach (Ahrens et al. 2003, 2004) to examine metaphors and Mapping Principles that use the source domain of BUILDING (*jianzhuwu)*. The corpus data will demonstrate that certain lexical mappings between the source-target domain pairings occur more frequently than other (Huang, Chung & Ahrens 2006). In addition, the corpus data will show the underlying reasons why different target domains select the same source domain of BUILDING to highlight different aspects of a building.

2 Using the Gigaword Corpus via the Chinese Word Sketch for the extraction of Mapping Principles from a source domain of BUILDING

Our goal is to establish the underlying reasons why the different target domains select the same source domain. We use the Gigaword Corpus and the Chinese Sketch Engish as tools to determine the underlying reasons.

The Chinese Sketch Engine (CSE, http://corpora.fi.muni.cz/chinese_all/) is a corpus processing system that was developed in 2005 (Huang et al. 2005) and was constructed by loading the Gigaword Corpus to the Sketch Engine

(Kilgarriff et al. 2005). The Gigaword Corpus contains about 1.12 billion Chinese characters, including 735 million characters from Taiwan's Central News Agency, and 380 million characters from China's Xinhua News Agency.

The Chinese Word Sketch can provide the information about a keyword's functional distribution (e.g. subject, object, etc), and the collocations in the corpus (i.e. the frequency a word collocates with a particular word). Furthermore, the Thesaurus can produce the synonym items that are automatically extracted based on common patterns of syntactic structures. For example, the six synonym candidates produced by the Thesaurus for the lexical item *jianzhuwu* "building" includes *jianwu* "a building", *fangwu* "a house", *zhuzhai* "a residence", *gongyu* "an apartment", *zhufang* "a residence", and *guozhai* "a house".

We use a nine-step paradigm in collecting and analyzing the corpus data from the Gigawords Corpus via the interface of the Chinese Sketch Engine. We will examine the metaphors using the source domain of BUILDING *jianzhuwu "building"* as an example and explain this paradigm in detail below.

First, we translate BUILDING into Chinese *jianzhuwu* "building". Second, in order to include all possible linguistic items for the concept of BUILDING in Mandarin Chinese, we select the synonym candidates automatically produced by the Thesaurus list for the keyword *jianzhuwu* "building" as we mentioned above.

Third, we use the Chinese Word Sketch to select 50 verbs that take these seven BUILDING lexical items as Subject and Object (i.e. items listed in the Subject_of and Object_of categories) and 25 lexical items as Possessor (i.e. items listed in the Possession category). For example, when *jianzhuwu* "building" functions as Subject, it collocates with the verb *daota* "to collapse"; when *jianzhuwu* "building" functions as Object, it collocates with the verb *laojiu* "old". In addition, when *jianzhuwu* "building" functions as Possessor, it collocates with *chuanghu* "a window". Finally, we collect 227 poential verbs and 139 potential nouns relating to the source domain of BUILDING.

This procedure is based on Ahrens' (2002) intuition-based approach to generate the mapping principles between target-source domain pairings. She proposed that each source domain, as a reflection of our real world knowledge, can be delimited with the three aspects: entities, qualities, and functions. Based on the same analogy, in order to limit our source domain of BUILDING, we only focus on lexical items that take the BUILDING words as Subject, Object and Possessor. These three syntactic categories can reflect the functions, quality and entities of the source domain of BUILDING.

Fourth, we use the Chinese Sketch Engine to distinguish the lexical items that specially relate to the BUILDING concept (i.e. *xinjian* "new built") from those that do not specially relate to BUILDING (i.e. *shen-gou* "to purchase"). In the Gigaword Corpus, the lexical item *shengou* "to purchase" not only collocates with a lexical token that relate to BUILDING, i.e. *guozhai* "a house", but also collocate with many tokens that relate to other knowledge domains, such as *ren* "people", *jijin* "a fund", *shenfenzheng* "an identification card", etc. We remove these ambiguous lexical items and select seventy-three verbs and twenty-six nouns that are associated with the BUILDING concept.

Fifth, we use the SUMO to define the abstract concepts that co-occur with these seventy-three verbs and twenty-six nouns. For example, SUMO defines the lexical word *guannian* "idea" as an abstract entity and the lexical item *duihua* "conversation" as a physical entity. SUMO can provide ontological nodes to indicate that whether a concept is *chouxiang* "abstract" or *wuzhi* "physical entity. The SUMO categorizes the lexis *guannian* "idea" in the domain of PROPOSITION *mingti* and its super-classes contain the abstract node. On the other hand, the lexical item *duihua* "conversation" is classified in the domain of COMMUNICATION *goutong* and its super-classes involve the physical node. The information of physical and abstract entities can help one distinguish whether the lexical words collocate with abstract concepts or not. We find eleven verbs (e.g. *chongjian* "to rebuild") and six nouns (i.e. *menchuang* "doors and windows") that collocate with an abstract entity.

Sixth, we define these abstract concepts with their target domain knowledge based on the WordNet senses and explanations. For example, the lexical items *ziliao* "data", *guannian* "ideas", *neirong* "contents", and *yijian* "an opinion" are identified as the target domain of IDEA because their WordNet senses and explanations include these words relating to the concept of proposition, such as "data", "concept", "facts", "idea", "opin-

ion", "belief" and "proof". This method allows one to define the target domain knowledge in an objective method and avoid determining the same domain simply based on native speakers' intuition.

Seventh, we classify the metaphoric instances based on the same target domains. For example, the metaphoric instances *laojiu* "aged", *cailiao* "a material", and *jichu* "foundation" are mapped to the lexical items *guannian* "an idea", *neirong* "contents", and *yijian* "an opinion", respectively. They are classified as the conceptual metaphor of IDEA IS A BUILDING because they have the same IDEA target domain. Eighth, we add up the frequencies of the lexical collocates of the same conceptual metaphor. For example, the metaphoric instance *laojiu* "aged" repeats within IDEA AS BUILDING because *laojiu* "aged" collocates 12 times with *ziliao* "data" and 7 times with *guannian* "an idea" in the Gigaword Corpus. Finally, we postulate mapping principles based on the most productive collocations.

3. Data Analysis, Results and Discussion

All the target-source domain pairings that have more than twenty instances are examined. In Table 1-7 below, the total number of metaphoric instances is given at the end of each table. Collocating frequency (i.e. Col. Freq.) indicates the number of tokens for each collocating lexical item in the Gigaword Corpus and the percentage refers to the percentage of the number of collocating items compared with the total number of metaphoric instances. The Mapping Principle for each conceptual metaphor is postulated according to the most productive collocations.

Table 1: AN IDEA IS A BUILDING

	Metaphor	Col. Freq.	%
Functions	*huisun* "to damage"	14	2.5
Qualities	*laojiu* "aged"	19	3.4
Entities	*jichu* "a foundation"	524	92.7
	chiliao "a material"	8	1.4
Total		565	100

The lexical items in the IDEA target domain include *guannian* "an idea", *neirong* "contents", *yijian* "an opinion", *lunzheng* "a proof", and *ziliao* "data". In the case of AN IDEA IS A BUILDING, the underlying reason has to do with foundations because the lexical item *jichu* "a foundation" has the highest collocating frequency (92.7%). The Mapping Principle postulated is: *An idea is understood as a building because buildings require bases for further build-up and an idea requires primary facts or evidence as bases for drawing a theory*. The corpus data demonstrates that concepts such as windows and doors, wall, construction, etc. are not mapped to IDEA.

Table 2: PRINCIPLES ARE BUILDINGS

	Metaphor	Col. Freq.	%
Function	*shousun* "to be destructed"	59	2.9
	chongjian "to reconstruct"	148	7.2
	jianshe "to construct"	126	6.1
Qualities	*laojiu* "aged"	5	0.2
Entities	*jichu* "a foundation"	1727	83.6
Total		2065	100

The lexical items in the PRINCIPLE target domain include *faling* "laws", *quanli* "right", *zhidu* "a system", *zhixu* "law and order", *yuanze* "principles", and *zhunze* "norms". In the case of PRINCIPLES ARE BUILDINGS, the underlying reason has to do with foundation because of the highest frequency of the collocating word *jichu* "a foundation" (83.6%). The Mapping Principle postulated is: *Principles are understood as buildings because buildings require bases for further build-up and principles require rules as bases for providing further guidance.*

When PRINCIPLES AS BUILDINGS are compared to IDEAS AS BUILDINGS, the concepts of *laojiu* "aged" and *jichu* "a foundation" are both mapped to the PRINCIPLES and IDEAS. However, some metaphoric instances are mapped to particular target domains. The lexical word *chongjian* "to reconstruct" and *jianshe* "to construct" are mapped to PRINCIPLES but not to IDEA.

Table 3: DIGNITY IS A BUILDING

	Metaphor	Col. Freq.	%
Functions	*shousun* "to be destructed"	60	68.2
	chonggjian "to reconstruct"	28	31.8
Total		88	100

The lexical items in the DIGNITY target domain include *zizun* "self-respect", *weixin* "prestige", *zunyan* "dignity" and *xinxin* "confidence". In the case of DIGNITY IS A BUILDING, the underlying reason has to relate to destruction because *shousun* "to be destructed" has the highest collocating frequency (62.8%). The Mapping Principle postulated is: *Dignity is understood as a building because buildings are destructed when a physical*

attack occurs and dignity is destructed when a verbal attack occurs.

Table 4 REPUTATION IS A BUILDING

	Metaphor	Col. Freq.	%
Functions	*shousun* "to be destructed"	319	90.0
	huisun "to destruct"	14	4.0
	sunhui "to damage"	7	2.0
	sunhuai "to damage"	7	2.0
	huihuai "to damage"	7	2.0
Total		354	100

The lexical items in the REPUTATION target domain include *mingsheng* "reputation", *mingyu* "fame", *xinyu* "prestige", *shangyu* "goodwill", *shengwang* "prestige", and *shengyu* "reputation". In the case of REPUTATION IS A BUILDING, the underlying reason has to do with destruction because *shousun* "to be destructed" is the most frequent collocation (90%). The Mapping Principle postulated is: *Reputation is understood as a building because buildings are destructed when a physical attack occurs and reputation is destructed when a verbal attack occurs.* The corpus data demonstrates that the words collocating with REPUTATION almost relate to the notion of destruction or damage.

Within the two conceptual metaphors, i.e. DIGNITY IS A BUILDING and REPUTATION IS A BUILDING, the BUILDING source domain repeatedly collocates with the abstract concepts of dignity and reputation. The Mapping Principles involve the concept of destruction. It seems that DIGNITY and REPUTATION are often expressed with more negative connotation.

Table 5: SPIRIT IS A BUILDING

	Metaphor	Col. Freq	%
Functions	*congjian* "to reconstruct"	577	97.5
Entities	*chuangfu* "a window"	9	1.5
	menchuang "doors and windows"	6	1.0
Total		592	100

The lexical item in the SPIRIT target domain only includes *xinling* "mind". In the case of SPIRIT IS A BUILDING, the underlying reason has to do with reconstruction because the lexical item *congjian* "to reconstruct" is the most prototypical lexical item (97.5%). The Mapping Principle postulated is: *Spirit is understood as a building because buildings are able to be reconstructed when building materials are ready and spirit is able to be reconstructed when the human emotion is ready to recover.* It is very interesting that the lexical items *chuangfu* "a window" and *menchuang* "doors and windows" are uniquely mapped to SPIRIT, but not mapped to other abstract domains we discussed previously.

Table 6: LIFE IS A BUILDING

	Metaphor	Col. Freq.	%
Functions	*chongjian* "to reconstruct"	274	100
Total		274	100

The lexical item in the LIFE target domain only includes *shenghua* "life". For LIFE IS A BUILDING, the underlying reason has to relate to reconstruction because *chongjian* "to reconstruct" is the only and highest collocation (100%). The Mapping Principle postulated is: *Life is understood as a building because buildings are able to be reconstructed when building materials are ready and life of a disadvantaged minority is able to be reconstructed when the social policies are established to help them.*

When SPIRIT IS A BUILDING is compared to LIFE IS A BUILDING, the source domain of BUILDING repeatedly collocates with the concepts of spirit and life. The two mapping principles involve the concept of reconstruction. It is likely that SPIRIT and LIFE are often expressed with more positive connotation.

Table 7: PROBLEMS ARE BUILDINGS

	Metaphor	Col. Freq.	%
Entities	*yaoshi* "a key"	31	66.0
	cailiao "a material"	16	34.0
Total		47	100

The lexical words in the PROBLEM target domain include *anjian* "a case", *wenti* "a question", *weiji* "a crisis", *aomi* "a mystery", *jiangju* "a deadlock", *mi* "a riddle", and *nanti* "a problem". For PROBLEMS ARE BUILDINGS, the underlying reason has to do with a key of a house because of the highest collocating frequency of *yaoshi* "a key" (66%). The Mapping Principle postulated is: *Problems are understood as buildings because buildings need keys for entering a house and problems need keys for solving the difficulty.* In addition, when PROBLEMS AS BUILDING is compared to other conceptual metaphors, the lexis "keys" is uniquely mapped to PROBLEM but never mapped to IDEA, DIGNITY, SPIRIT, etc.

Table 14 shows the underlying reasons that different target domains select the same source domain of BUILDING. These underlying reasons can be framed at a linguistic level based on the analysis of the conventional mappings between the source and target domain pairings (Ahrens 2002).

Table 8: Four aspects of a building used for different source domains

Aspects of a building	Source domains
Foundations	IDEAS, PRINCIPLES
Destruction	DIGNITY, REPUTATION
Reconstruction	SPIRIT, LIFE
Keys	PROBLEMS

The contrastive linguistic analysis shows why IDEA is discussed metaphorically in terms of BUILDING for borrowing the aspect of foundation while PROBLEMS is discussed metaphorically in terms of BUILDING for borrowing the notion of a key. The emphasis for IDEAS is on foundations in order to indicate an initial stage of something. For example, when people talk about IDEAS, they have to take facts as supporting evidence in advance to draw a new theory. Foundations act as a primary stage to do something else. On the other hand, the reason to emphasize the dimension of keys of a house for the abstract concepts, e.g. PROBLEMS, CRISIS, and DEADLOCK rather than foundations because foundations are no more an important issue when people talk about PROBLEM. Instead, it is critical to solve a problem with the necessary person, object, etc.

Finally, the corpus data demonstrates how people use metaphors in daily discourse. For example, the concepts of reputation and dignity are frequently discussed negatively. They borrow the aspect of destruction from the BUILDING source domain having to do with "damage" to express the negative sense. On the other hand, the concepts of spirit and life are frequently discussed positively. They borrow the aspect of construction from the BUILIDING source domain having to do with "reconstruction" to convey the positive sense.

4 Conclusion

In this study, we use a large-scale of corpus, i.e. the Gigaword Corpus in combination with the Chinese Sketch Engine, to examine the underlying reasons between source and target domain pairings. We employ a frequency-based collocational approach to analyze conceptual metaphors with the source domain of BUILDING, and determine their Mapping Principles. The corpus analysis verifies that the underlying reasons between source and target domain pairings can be extracted based on the most productive collocation. For each conceptual metaphor, we can find out that a particular lexical mapping occurs more frequently than the others. Second, the corpus data demonstrates the underlying reasons why the different target domains select the same source domains. In particular, we find out that the same source domain BUILDING is repeatedly mapped to nine different target domains to highlight four different dimensions of a building. The concept of foundation collocates with IDEAS. The concept of a key collocates with PROBLEMS. The notion of destruction collocates with REPUTATION and the notion of reconstruction collocates with SPIRIT. Finally, we also find out that the different target domains select the same BUILDING source domain for the same underlying reasons. For example, the concepts of ideas and principles select the aspect of foundations as the underlying reason from the BUILDING source domain to emphasize an initial stage of something from which further advances can be made.

To conclude, this corpus-based study follows the Conceptual Mapping Model's proposal that the lexical mappings can be acquired through a contrastive linguistic analysis. Further research will employ the same method discussed herein to explore more metaphors in Mandarin Chinese that use the other source domains, such as FOOD, GAME, FIRE, etc. It is hoped in this way that conceptual mappings will no longer be considered ad hoc results of source-target domain pairings but instead involve principled explanations based on prototypical mappings.

References

Ahrens, Kathleen. 2002. "When Love is not Digested: Underlying Reasons for Source to Target Domain Pairing in the Contemporary Theory of Metaphor". In YuChau E. Hsiao (ed.) In the *Proceedings of the First Cognitive Linguistics Conference,* Taipei: Cheng-Chi University, 273-302.

Ahrens, Kathleen, Siaw-Fong Chung, and Chu-Ren Huang. 2003. "Conceptual Metaphors: Ontology-based Representation and Corpora Driven Mapping Principles." In the *Proceedings of the ACL Workshop on the Lexicon and Figurative Language,* 35-41.

Ahrens, Kathleen, Siaw-Fong Chung and Chu-Ren Huang. 2004. "From Lexical Semantics to Conceptual Metaphors: Mapping Principle Verification with WordNet and SUMO." In Ji Donghong, Lua Kim Teng, and Wang Hui (Eds). *Recent Advancement in Chinese Lexical Semantics: Proceedings of 5th Chinese Lexical Semantics Workshop (CLSW-5)*. Singapore: COLIPS, 99-106.

Huang, Chu-Ren, Adam Kilgarriff, Yicing Wu, Chih-Min Chiu, Simon Smith, Pavel Rychly, Ming-Hong Bai, and Keh-Jiann Chen. 2005. "Chinese Sketch Engine and the Extraction of Collocations." In the *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, October 14-15. Jeju, Korea, 48-55.

Huang, Chu-Ren, Siaw-Fong Chung and Kathleen Ahrens. "An Ontology-based Exploration of Knowledge Systems for Metaphor". 2006. In K. Rajiv, R. Ramesh, & R. Sharman (Eds.), *Ontologies: A Handbook of Principles, Concepts and Applications in Information Systems.* Volume 14. Springer.

Kilgarriff, Adam, Chu-Ren Huang, Pavel Rychly, Simon Smith, and David Tugwell. 2005. *Chinese Word Sketches.* ASIALEX 2005: Words in Asian Cultural Context. Singapore.