Using Chinese Gigaword Corpus and Chinese Word Sketch in linguistic research

Jia-Fei Hong, Chu-Ren Huang

National Taiwan University, Graduate Institute of Linguistics,
No. 1, Sec. 4, Roosevelt Road, Taipei, 106 Taiwan
Institute of Linguistics 128, Section 2,
Academia Road 115, Taipei, Taiwan, R.O.C

{jiafei, churen}@gate.sinica.edu.tw

Abstract. We explore the possibility of deeper linguistic research based on corpus and computational linguistic tools in this paper. In particular, we adopt Chinese Word Sketch, the application of Word Sketch Engine to Chinese GigaWord Corpus, for linguistic research. We apply Chinese Sketch Engine results to deeper linguistic account such as selectional restriction and event type selection. The study is based on the comparison of two basic verbs of ingestion: chi1 'to eat' and he1 'to drink'.

Key words: Chinese Word Sketch, selectional restriction, event type, Corpus-based linguistic research

1 Introduction

In this paper, we explore the potential of Chinese Word Sketch (CWS) as a tool for deeper linguistic research. The CWS is a combination of the Chinese GigaWord Corpus (Huang et al. 2005) with the linguistic search tool of Word Sketch Engine (Kilgarriff et al. 2004). It is well documented that the Sketch Engine is a very powerful tool for extracting meaning grammatical relations given a sufficiently large corpus (Kilgarriff et al. 2004, Huang et al. 2005). We show in this paper how CWS can be applied to confirm and refine human lexical semantic work, as well as be applied to deeper linguistic account.

The set of linguistic fact that we will account for involve the two basic verbs of ingestion chi1 'to eat' and he1 'to drink'. We draw directly from the corpus-based manual analyses from Chinese Wordnet Group (CWN Group) to define the sense divisions for "chi1" and "he1". Chinese Word Sketch Engine

helps us to verify and refine the manual analysis, especially with the event type and argument roles for “chi1” and “he1”.

In addition, we are concerned with the semantic attributes shared or distributed among these argument roles. We further explore the linguistic accounts of *selectional restriction and event type selection* based on these distributions.

The remaining part of this paper starts with an introduction to the GigaWord Corpus and Word Sketch Engine. We next explicate the scope and and goals of the current study. Thirdly, we detail the collocation data, as well as how they support and refine manual sense analysis. Fourthly, we show how the derived generalizations can be applied in deep linguistic account. Finally, we conclude with predictions and future work.

2 GigaWord Corpus

The Chinese Gigaword Corpus contains about 1.1 billion Chinese characters, including more than 700 million characters from Taiwan’s Central News Agency, and nearly 400 million characters from China’s Xinhua News Agency. Before loading Chinese Gigaword into Sketch Engine, all the simplified characters were converted into traditional characters, and the texts were segmented and POS tagged using the Academia Sinica segmentation and tagging system (Huang et al., 1997). The segmentation and tagging was performed automatically with automatic and partially manual post-checking. The precision accuracy is estimated to be over 95% (Ma and Huang 2006).

3 Chinese Word Sketch

The two challenges to corpus-based computational approaches to linguistic analysis are to acquire enough data to show linguistic distribution, and to design efficient tools for extracting linguistically significant generalizations from vast amount of data. Kilgarriff et al. (2004) developed the Sketch Engine to facilitate efficient use of gargantuan corpora. The Sketch Engine (SKE, also known as the Word Sketch Engine) is a novel Corpus Query System incorporating word sketches, grammatical relations, and a distributional thesaurus.

The advantage of using the Sketch Engine as a query tool is that it pays attention to the grammatical context of a word, instead of just pouring an arbitrary number of adjacent words. In order to show the cross-lingual robustness of the Sketch Engine as well as to propose a powerful tool for collocation extraction based on a large scale corpus with minimal pre-processing; we constructed Chinese Word Sketch Engine (CWS) by loading the Chinese Gigaword to the Sketch Engine (Kilgarriff et al., 2005). All components of the

Sketch Engine were implemented, including *Concordance, Word Sketch, Thesaurus and Sketch Difference*.

4 Motivation and Goals

Studies of the synonyms of from the same semantic field, such as Tsai et al. (1998) and Huang and Hong(2005) established that a set of syntactic and distributional variations may be accounted for by one fundamental semantic contrast. Their study of Sheng1 and Yin2 (both 'sound') showed the grammatical difference between the two can be accounted for by their contrast in lexical semantics, where sheng1 focuses on production and yin1 focuses on perception. This accounts for many contrasts for the two near synonyms, including the following contrast.

(1a) 多田忽然**放聲/*放音**大笑，用生硬的德文，一再向我道歉。
Duo1 tian2 hu1 ran2 **fang4 sheng1/* fang4 yin1** da4 xiao4, yong4 sheng1 ying4 de5 de2 wen2, yi2 zai4 xiang4 wo3 dao4 qian4.
Duo tian suddenly to laugh aloud, use awkward German, again and again to I apologize.
'Duo-tian laughs aloud suddenly and apologizes to me again and again in awkward German'

(1b) 他們只能靠上課**錄音/*錄聲**和班上同學的幫忙，學習上格外辛苦。
Ta1 men5 zhi3 neng2 kao4 shang4 ke4 **lu4 yin1/* lu4 sheng1** han4 ban1 shang4 tong2 xue2 de5 bang1 mang2, xue2 xi2 shang4 ge2 wai4 xin1 ku3.
'They can only depend on recordings of classes and helps from classmates, therefore It has been extremely difficult for them to study."

The near synonym pair sheng1 and yin1 has near complimentary distributions to clearly indicate their semantic contrast. In this study, we will look at the verbs of ingestion chi1 'to eat' and he1 'to drink'. What is interesting for chi1 and he2 are two folded. First, they have overlapping meaning where one is an elaboration of the other, called troponym by WordNet; second, they have very rich semantic extensions. We apply CWS to explore the possibility of constructing an account of the complex meaning extensions based on automatically analyzed distributional data from CWS.

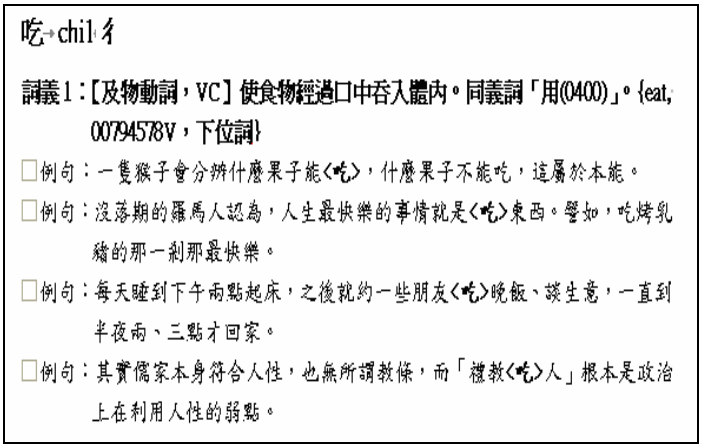
5 Data Collection and Sense Analysis

Date from the smaller (5 million words) Sinica Corpus shows clearly that

the dominant sense of chi1 is to ingest solid food, while he1 is to ingest liquid. Based on these data and detailed manual analysis, the sense inventory of chi1 and he1 are given below (Huang et al. 2006).

5.1 Original Sense and Selectional Restriction

The CWN Group gives the original sense for chi1 and he1 as below:


吃→chi1ㄔ

詞義1：[及物動詞，VC] 使食物經過口中吞入體內。同義詞「用(0400)」。{eat, 00794578V，下位詞}

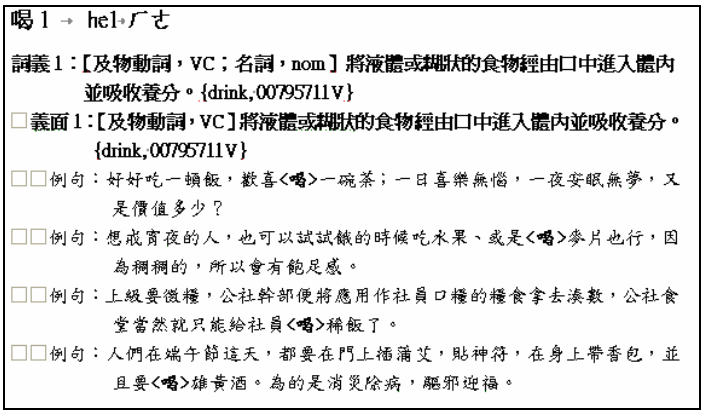
例句：一隻猴子會分辨什麼果子能<吃>，什麼果子不能吃，這屬於本能。

例句：沒落期的羅馬人認為，人生最快樂的事情就是<吃>東西。譬如，吃烤乳豬的那一剎那最快樂。

例句：每天睡到下午兩點起床，之後就約一些朋友<吃>晚飯、談生意，一直到半夜兩、三點才回家。

例句：其實儒家本身符合人性，也無所謂教條，而「禮教<吃>人」根本是政治上在利用人性的弱點。

Table1: Original sense of “chi1”


喝 1 → he1ㄏㄜ

詞義1：[及物動詞，VC；名詞，nom] 將液體或糊狀的食物經由口中進入體內並吸收養分。{drink, 00795711V}

義面1：[及物動詞，VC] 將液體或糊狀的食物經由口中進入體內並吸收養分。{drink, 00795711V}

例句：好好吃一頓飯，歡喜<喝>一碗茶；一日喜樂無惱，一夜安眠無夢，又是價值多少？

例句：想成宵夜的人，也可以試試餓的時候吃水果、或是<喝>參片也行，因為稠稠的，所以會有飽足感。

例句：上級要徵糧，公社幹部便將應用作社員口糧的糧食拿去湊數，公社食堂當然就只能給社員<喝>稀飯了。

例句：人們在端午節這天，都要在門上插蒲艾，貼神符，在身上帶香包，並且要<喝>雄黃酒。為的是消災除病，驅邪迎福。

Table2: Original sense of “chi1”

Table1 and table2, show that the two verbs of ingestion differ in that “chi1” takes solid object(e.g. fruit, food, dinner and so on), while he1 takes liquid objects (e.g. tea, wine and so on). This should offer a classical case of selectional restriction. However, corpus data show that there are many significant counter-examples, such as “chi1 nai3 shui3” (to eat milk) or “he1 xi1 fan4” (to drink porridge). This set of data is challenging for a simple selectional restriction account.

5.2 Extended Sense and Metaphorical Sense

CWN sense lexicon also extended senses and metaphorical sense differ two verbs such as below:

詞義 8：【及物動詞，VC】比喻遭受後述事物的攻擊。同義詞「挨2(0121)」。{suffer, 01444459V}
□例句：他<吃>了三顆子彈，兩顆在前胸，一顆在下腹。
□例句：蘇星用石子擊中了行刑兵的手腕，讓俾劍免<吃>了一頓鞭。

詞義 9：【及物動詞，VC】比喻經歷後述負面事件。{suffer, 01444459V}
□例句：說得誇張一點，過去一年來，<吃>官司的銀行董事長或總經理，多到兩隻手都數不過來。
□例句：中華成棒赴澳移地訓練，昨第一次與澳洲國家對交手，結果澳洲怪投，使中華<吃>苦頭，以 0：9 慘敗。
□例句：省體在這場戰役，特別將亞運國手全部安排下場，選連<吃>敗仗，下半場七十分鐘，銘傳的突擊小組再次出發。
□例句：台北社區・建設局長巡視，萬客隆，內湖店，<吃>閉門羹，該店迄今仍是違規營業，但店方表示已申請為批發專用區。

詞義 10：【及物動詞，VC】比喻佔便宜。{flirt, 00330518N}
□例句：劉若英被<吃>豆腐、過色狼的經驗太多了，公司只好請私人保鏢來保護她。
□例句：因為沙灘上幾乎所有的女郎，全部都穿著比基尼泳衣，在球員前面逛來逛去，讓眼睛大<吃>冰淇淋。

Table3: Extended senses of “chi1”

The above analysis suggests that one clear clue for a meaning extended or metaphorical use is the violation of selectional restriction. This is, in turn, facts that should be extendable from CWS.

5.3 Neutralized Selectional Restrictions

One set of challenging facts for selectional restrictions involves cases where they are neutralized. For example, when an object has both solid and liquid attributes, object will be selected both by “chi1” and “he1” such as below:

(2) chi1 xi1 fan4
to eat porridge

(3) he1 xi1 fan4
to drink porridge

These neutralization effects can also be found with metaphoric uses. For instance, both wedding banquet (xi3 jiu3) and afternoon tea time (xia4 wu3 cha2), can be selected by both verbs “chi1” and “he1”.

6 Word Sketch Patterns

One of the most powerful functions of the Word Sketch Engine is the Sketch Difference. This is a very efficient tool for doing contrastive studies. It compares the salient collocating grammatical relations of the two keywords and returns grammatical relations that are 1) shared by both keywords, 2) unique to only one keyword, or 3) that are highly salient for one but not salient for the other. When we examine the sketch difference for “chi1” and “he1”, in terms of their objects, the significant results are shown as from table 6 to table 8.

Common patterns

	More usage for「chi1」	Common usage	More usage for「he1」
object	yao4(medicine)、dong1 xi1(foodstuff)、xi1 fan4(rice porridge)…	nai3(milk)、zhou1(porridge)、leng3 yin3(cooling drink)、nai3 shui3 (milk)…	jiu3(wine)、xi3 jiu3 (wedding banquet)、niu2 nai3 (milk)、ku3 shui3(complaints)…

Table4: The common patterns for "chi1" and "he1"

Only patterns

	"chi1" only patterns
object	yao4(medication)、fan4(dinner)、shui3 guo3(fruit)、zao3 can1(breakfast)；*luo2 si1*(screw)、*nai3 zui3*(pacifier)、*zi3 dan4*(cartridge)、*qiang1 zi5*(firearm)；**kui1**(loss)、**bai4 zhang4**(a lost battle)、**lao3 ben3**(original capital)、**hong2 pai2**(red card)；***xi1 fan4***(porridge)、***xi3 jiu3***(wedding banquet)、**bi4 men2 geng1**(to slam the door in one's face)…

Table5: The only patterns for "chi1"

	"he1" only patterns
object	jiu3(wine)、cha2(tea)、ka1 fei1(coffee)、kai1 shui3(water)；*ku3 shui3*(complaint)；***xi1 fan4***(porridge)、***xi3 jiu3***(wedding banquet)；**nai3 fen3**(milk powder)、**yan4 wo1**(edible nest of cliff swallows)…

Table6: The only patterns for "he1"

The classical theory of selectional restriction stipulates that verb checks certain semantic features in an argument and allows only those whose features fit the selection. We have followed this theory and assumed that the features [+/- food], [+/- solid], [+/- liquid] were at work. However, the collocational patterns extracted from Chinese Word Sketch strongly suggests otherwise. It is clear that meaning plays a central role in these selections. However, the case of neutralized selection as well as metaphorical and metonymic extensions point to the inadequacy of a feature checking account.

We propose that a more appropriate theory to account for the semantic selection facts are concept rather the feature based. In particular, we propose that the rich knowledge structure of an ontology is more appropriate for semantic content and selection. We will take three examples of neutralized selection so show how a more complete representation of conceptual location offers a better

account.

First, take xi1fan4 porridge for example. There is no doubt that porridge is a kind of food and it will be so assigned to this ontological node. However, our lexical knowledge of porridge also include that it contains two important ingredients: rice (solid) in soup (liquid). Conventionally, it is a kind of rice diluted by water (xi1-fan4, where fan is rice) by lexical combination. Hence in a merged lexical ontology, it should inherit properties from both solid and liquid materials, depending on whether the focus on the rice or the soup. This kind of representation not only account for the fact that both verbs are allowed, but also picks up the subtle focus.

Second, we account for the metonymical extension case involving chi1/he1 xi3jiu3. Note that xi3jiu3 'happy-wine, wedding banquet' itself is a metonymical extension, using the wine/liquor drunk at the wedding banquet to refer to the event. A wedding banquet is a sub-type of banquet, where both eating and drinking are the most salient activities. In other words, xi3jiu3 will be assigned a conceptual location of being a banquet. Since it by form is a kind of food (due to the lexical head of 'jiu'), both verbs for ingestion is coerced. However, since the activity involves both eating and drinking, both verbs were allowed.

Lastly, in the case of 'chi1 nai3 shui3' and 'he1nai3shui3', it is interesting to note that he1 nai3 shui3 'literally, to drink (x's) milk' is ambiguous between the literal and metaphoric reading 'to be raised by, or with X ideology'. However, chi1 nai3hui3 'literally, 'to eat someone's milk' only allows the metaphoric reading. Again, there is no doubt the milk is liquid food in the literal interpretation. However, when metaphoric uses are involved, nai3sui3 refers to nourishment for either the body or the soul. In this case, again, a speaker chose to use both verbs of ingestion. This is because the metaphor involves nurture and nourishment, which typically involves feeding.

To sum up, we show that the richer conceptual representation of ontology allows us not both describe and account for novel linguistic facts more succinctly.

7 Conclusion

In this paper, we show that the Chinese Sketch Engine is a very powerful tool for deeper linguistic analysis. While examine the keywords chi1 and he1, we were able to propose an improved account of selectional restriction based on the distributional data We were also able to establish a model of event type selection, where we can predict the meaning of a non-typical event type object, as well as predict metaphoric meaning.

Using this search engine could help us to find out the rough sense divisions.

And then we take the man-made analyses from CWN to show that the results are accurate by the semiautomatic research of Chinese Word Sketch Engine.

Reference

1. Huang, Chu-Ren, Adam Kilgarriff, Yicing Wu, Chih-Min Chiu, Simon Smith, Pavel Rychlý, Ming-Hong Bai, and Keh-Jiann Chen. 2005. Chinese Sketch Engine and the Extraction of Collocations. In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, 48-55. October 14-15. Jeju, Korea.
2. Huang Chu-Ren, Jia-Fei Hong. 2005. Deriving Conceptual Structures Form Sense: A Study of Near Synonymous Sensation Verbs. Journal of Chinese Language and Computing (JCLC). Volume 15, No 3, 2005, Singapore.
3. Kilgarriff, Adam, Chu-Ren Huang, Pavel Rychlý, Simon Smith, and David Tugwell. 2005. Chinese Word Sketches. ASIALEX 2005: Words in Asian Cultural Context. Singapore.
4. Ma Wei-yun, and Chu-Ren Huang. 2006. Uniform and Effective Tagging of a Heterogeneous Giga-word Corpus. Presented at the 5th International Conference on Language Resources and Evaluation (LREC2006). Genoa, Itlay. 24-28 May, 2006.
5. Tsai, Mei-Chih, Chu-Ren Huang, Keh-Jiann Chen, and Kathleen Ahrens. 1998. Towards a Representation of Verbal Semantics--An Approach Based on Near Synonyms. *Computational Linguistics and Chinese Language Processing.* 3(1): 61-74.
6. 黃居仁 主編。洪嘉馡, 陳韻竹. 2006. 執行總編輯。《意義與詞義》系列《中文詞彙意義的區辨與描述原則》. 中央研究院語言所詞網小組與詞庫小組技術報告.

Website Resource

1. Sinica Corpus. http://www.sinica.edu.tw/SinicaCorpus/
2. Sketch Engine (Chinese) http://corpora.fi.muni.cz/chinese_all/
3. Sketch Engine (English) http://www.sketchengine.co.uk/