CSCanada
Studies in Literature and Language
Vol. 8, No. 1, 2014, pp. 48-56
DOI:10.3968/j.sll.1923156320140801.4192

ISSN 1923-1555[Print]
ISSN 1923-1563[Online]
www.cscanada.net
www.cscanada.org

How Do People Process Ambiguous Strings

Liu, Hsiu-Ying[a]; Kuo, Cheng-Chung[b],*

[a]Associate professor. Department of Foreign Languages and Literature, Asia University, Taichung, Taiwan, China.
[b]Department of Cultural and Creative Industries, National Taichung University of Education, Taichung, Taiwan, China.
*Corresponding author.

Received 13 November 2013; accepted 5 February 2014

Abstract

This article combines ambiguity phenomenon with Chinese word segmentation to observe how human being conduct language processing to clarify ambiguity between overlapping ambiguity and combination ambiguity. Artificial intelligence will easily missegment these two strings, while the study tries to introduce optimality theory to discover possible base of these two types of ambiguity comprehended by general people. According to the result, the key to clarify ambiguity is context, idiomaticity and word frequency.

Key words: Ambiguity; Chinese word segmentation; Language processing; Compound

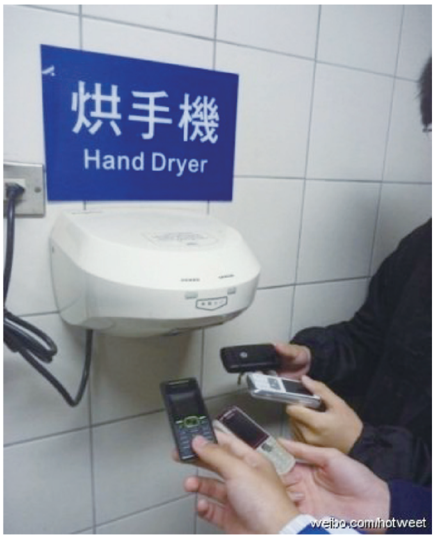
Liu, Hsiu-Ying, Kuo, Cheng-Chung (2014). How Do People Process Ambiguous Strings. *Studies in Literature and Language, 8*(1), 48-56. Available from: http://www.cscanada.net/index.php/sll/article/view/j.sll.1923156320140801.4192
DOI: http://dx.doi.org/10.3968/j.sll.1923156320140801.4192

INTRODUCTION

The following photo about Mainlanders at public toilet in Taiwan is spread on the internet.

Mainlanders disregarded the English translations but interpreted the signage as "cellphone dryer". Although such thing would not happen between Taiwanese people, we would actually make fool of ourselves in life in a similar way.

Ambiguity exists in language and normally causes misunderstanding and jokes. There are many causes of ambiguity phenomenon including phonetic, vocabulary or grammar. There are also many ambiguity phenomena in Chinese, in which grammatical structure is the most interesting one. In terms of written form, Chinese is piled up by each single Chinese character and may lead to ambiguity easily when there is no punctuation. There will be a conversation with Charlotte & Jaden and lead to communication failure if both parties choose different meanings to interact with.



http://eznewlife.com/?p=6243

Figure 1
xxxxxx

How do most people understand each other while facing an ambiguity? Do they have a set of criteria to deal with the ambiguity phenomenon? The study will try to observe the tendency of people's choice from the perspective of ambiguity phenomenon of Chinese. We will invite 300 students to conduct semantic selection upon 10 ambiguous sentences, with discussion and analysis on the outcomes conducted afterwards, to expect to find out the principle for ambiguity management. According to initial observation, it seems that interviewees follow a set of rules while dealing with ambiguity. The results of this study not only show rules of language management of people, but also provide as a reference to natural language processing managed by artificial intelligence as well as Chinese learning and teaching.

1. AMBIGUITY

Ambiguity refers to the phenomenon that a linguistic unit (up to a sentence, down to a word) has more than one interpretation. According to many studies (Lü, 1984; Zhao 1992; Tsui, 2008; Zhang, 2009), ambiguity in Chinese can be phonetic, lexical or grammatical. Phonetic ambiguity is mainly caused by homonymy, lexcial ambiguity is always brought by polysemous words while different structural hierarchies, structural relations or semantic relations will result in grammatical ambiguity.

The ambiguity phenomenon explored by this article is mainly caused by the following two types of strings:

(1) The string that can be analyzed as either a phrase or a compound
e.g., *xiao-ren*
(2) The string whose middle element could combine with either the preceding element to form a phrase or the following element to form a phrase or compound
e.g., *an-quan xing xing-wei*

Basically, these two phenomena are grammatical ambiguity caused by different grammatical structures. In addition, they both are associated with compounds.

1.1 Studies on Ambiguity

There are many studies coping with Chinese ambiguity. Generally, ambiguity in Chinese can be divided into three types: phonetic ambiguity, lexical ambiguity and grammatical ambiguity. According to Chen (2010), phonetic ambiguity appears mostly in colloquial language. It happens because a syllable might correspond to multiple morphemes in Chinese.There are three types of phonetic ambiguity.

1. ambiguity caused by **homonym**
A: Qingwen nin guixing
"What's your family name?"
B: "Zhang."(would be written as two or more Chinese characters)
2. ambiguity caused by **polyphony**
Zhege ren hao shuohua
this man speak
a) Zhege ren hao53 shuohua "The man likes to talk (talkative)."
b) Zhege ren hao213 shuohua "The man is easy to negotiate with."
3. ambiguity caused by **stress**
Zuihao mai yi-ge
would better buy one
a) Zui hao mái yige
"You'd better buy one (rather than borrow one from others)."
b) Zui hao mai yíge
"You'd better buy one (but not two or more)."

Lexical ambiguity appears when a word has more than one possible meaning, briefly divided into three types:

1. ambiguity caused by **polysemy**
Ta yijing zou le
he already go ASP
a) He passed away.
b) He has left.
2. ambiguity caused by **homonym**
Zheer de ren duoban shi daxuesheng
here people be college student
a) People here are mostly college students.
b) People here might be college students.
3. ambiguity caused by **homograph**
Ta can jun qu le
he army go ASP
a) He joined the army.
b) He reviewed the troops.

Grammatical ambiguity refers to ambiguity explained by differences in syntax. It is usually brought by different grammatical relations or syntactic structures. There are four possible factors.

1. ambiguity caused by **multi-category word**
Chouti meiyou suo
drawer no lock
a) The drawer is unlocked.
b) There is no lock for the drawer.
2. ambiguity caused by different structures
Zhezhang zhaopianli shi xiaoming han xiaogang de baba
this sheet photo in be and GEN father
a) Xiao-ming and Xiao-gang's father is in the photo.
b) Xiao-gang's father and Xiao-ming are in the photo.
3. ambiguity cause by different possessions
Zhe shi xiaoming de zhaopian
this be GEN photo
a) This is a photo of Xiao-ming. (Xiao-ming is in the photo.)
b) This photo is possessed by Xiao-ming.
4. ambiguity caused by unclear referent
Ta gang zhuan dao zhesuo xuexiao henduo ren dou bu renshi
he just transfer to this school many people all not know
a) He has just transferred to this school, so many people don't know him.
b) He has just transferred to this school, so he knows very few people.

Basically, the two strings to be studied in this present research would bring about grammatical ambiguities caused by different structures. Nonetheless, such strings are not mentioned in previous studies, and there is no discussion about how human beings process these ambiguous strings.

Compound vs. Phrase

Compounding is a popular way for word formation in Chinese, and the prevalence of compounds increases the importance of it. According to Li and Thompson (1981, p.47), Chinese compounds could be grouped into three types semantically.

a. There may be no apparent semantic connection between the meaning of the compound and the meaning of its constituents, e.g., *fengliu* "amorous".

b. There may be a metaphorical, figurative, or inferential connection between the meaning of the compound and the meanings of its component parts, e.g., *maodun* “contradictory”.
c. The meaning of the compound may be directly related or identical to the meanings of its components, e.g., *xizao* “take a bath”.

A compound is composed of two or more free morphemes. Based upon the internal structure, in Chinese, there are five types of compounds.

1. subject-predicate: xin-teng (to feel anguished)
2. predicate-object: shou-jiu (to stick to old ways)
3. modifier-head: xiao-ren (a villain)
4. predicate-complement: jian-shao (to decrease)
5. parallel/coordinate: kai-guan (a switch)

A phrase is a group of words that forms a **constituent** and functions as a single unit in a sentence. It is composed of two or more words. A phrase is different from a compound in that the combination does not result in a single word. Interestingly, the structures of phrases could be grouped into the five types the same as those of compounds:

1. subject-predicate: huodong jieshu (The activity ends.)
2. predicate-object: kao yanjiusuo(to take the entrance exam of graduate school)
3. modifier-head: hen hao (very good)
4. predicate-complement: da si (beat someone to death)
5. parallel/coordinate: fu mu (father and mother)

Since compounds and phrases might share the same structure, it is supposed that ambiguity arises easily and people would encounter difficulty in processing the strings, especially when the string could be segmented as both a compound and a phrase. *Xiao-ren* could be realized either as a compound referring to a villain or a phrase indicating a tiny person[1].

2. LANGUAGE PROCESSING

2.1 Artificial Intelligence

Generally speaking, it mostly refers to processing of natural language on machine when it comes to language processing. Therefore, most of the studies are conducted by experts/scholars of computational linguistics, some by eye tracker, and some by statistics.

Chinese Word Segmentation

Chinese word segmentation has been a big issue for natural language processing by artificial intelligence. Chinese word segmentation is to segment a sequence of Chinese characters into separate words. It is a process to re-assemble a sequence of characters upon given regulation. Words appear to be the basic unit in English. English separates words by space so that computers or artificial intelligence can easily determine whether a string is a word or not. Differently, the basic unit of Chinese is Chinese character. Since there is no space between Chinese characters, artificial language analyzer will mis-segment a Chinese string easily. What is interesting is that the most common error strings are exactly the same as the strings mentioned in the study.

According to previous studies, computers always make mistakes while segmenting the following two strings:

A) overlap ambiguity string (OAS)

Take a character string ABC as an example. If both AB and BC are words, the segmentation of ABC can be either AB/C or A/BC depending on different context. The ABC is called an overlap ambiguity string (OAS).

e.g., zhu he cheng
a) Women/ xiaozu / hecheng / qingqi le
our group synthesize hydrogen
“Our group synthesized hydrogen.”
b) zuhe / cheng / fenzi
“to make up to be a molecule”

B. combination ambiguity string (CAS)

Take a character string AB as an example. If A, B, and AB are words, the segmentation of AB can be either A/B or AB. The AB is called a combination ambiguity string (CAS).

e.g., ma shang
a) Ta/ cong/ **ma/ shang**/ xia/ lai
he from horse ups down come
“He got down from the horseback.”
b) Wo/ **mashang**/ jiu/ lai/ le
I immediately will come ASP
“I will come immediately.”

To artificial intelligence, there are three main types of segmentation algorithms at present:

1. Word segmentation algorithm based on string matching
2. Word segmentation algorithm based on comprehension
3. Word segmentation algorithm based on statistics

The first one, also called maximum matching method, is dictionary-based and the most commonly used. It conducts the matching between Chinese strings to be analyzed and entries in a fully big machine dictionary. A term is identified out of a successful matching if it is found in the dictionary. This method is easy to realize and reports speedy segmentation, but it fails to correctly deal with combination ambiguity strings and some complex overlapping ambiguity ones. The second type is mainly a process similar to people’s comprehension, which conducts grammar and semantic analysis and uses grammatical information and semantic information to deal with ambiguity phenomenon while handling word segmentation. Such a method requires bulky language knowledge and information. However, since linguistic information is basically hard to be organized as a format to be read by the machine directly, the method is far from completion and still under experimental stage. The third type of the methods is a statistical-based approach which

[1] In Chinese, there is a way to test whether a string is a phrase or not. A phrase could transform to a coordinating structure. If *xiao ren* is a phrase, it can transform to *you xiao you he de ren* (a person who is small and black).

is proclaimed as non-dictionary word segmentation or statistical word segmentation method. It is also called word frequency method in that this algorithm counts the frequency of collocation of two or more adjacent characters in the data to get the co-occurrence information of the characters. The information would further provide the closeness of combination between Chinese characters. The closer the characters, the more possible they form a word. This method, nonetheless, extracts not only words. It usually selects some common phrases that report high concurrent frequency, such as *wo-de* 我的 “mine” and *xu-duo-de* “many”.

2.2 Garden Path Sentence

GARDEN PATH refers to the saying “to be led down the garden path”, meaning “to be misled”. In psycholinguistics, a garden path sentence is temporarily ambiguous or confusing because it contains a word group which appears to be compatible with more than one structural analysis. Gardenpath sentences are easily misunderstood (they lead you down the garden path) even though they are all grammatical.

A garden path sentence is different from an ambiguous sentence in that the former has only one correct interpretation whereas the latter allows two or more interpretations. Garden path sentences are used mainly to illustrate the fact that human beings process language one word at a time when they read. To parse a sentence, human beings always adopt a serial mechanism, word by word from left to right. Such a parsing way might bring about improper parsing that turns out to be a dead end. A second parsing should then be required to get the correct interpretation.

a. Peter knew the answer immediately.
b. Peter knew the answer would be false.

As for sentence (b), a reader usually starts to parse this as an ordinary active transitive sentence but stumbles when reaching the word *would*. At this point, the reader is forced to backtrack and look for other possible structures. It may take some rereading to realize that *the answer* is in fact a part of a subordinate clause, implying that *the answer* is the subject of the embedded clause. The correct reading is then: “Peter – knew- (that) the answer would be false.”

As to the following sentences, (c) is a garden path sentence and (d) is a sentence with ambiguity.

c. Da hua chi chi de deng le yi nian
d. Da hua chi deng le yi nian

The reader might first parse *da-hua-chi* as a noun phrase functioning as the subject of the sentence. However, he/she stumbles when reaching the other *chi*. The second parsing is required to make *da-hua* the subject and *chi-chi-de* an adverb, resulting “*Dahua-chichide-dengleyinian*.” Sentence (d) has two possible interpretations.

(1) Dahua/ chideng le yinian
Dahua crazy wait ASP one year
“Dahua has crazily waited for a year.”
Da huachi deng le yinian
big anthomaniac wait ASP one year
“The big anthomaniac has waited for a year.”

Although there are differences between garden path sentences and ambiguous sentences, they are not totally irrelevant. Garden Path sentences normally have local ambiguity; an ambiguous sentence would undergo the second parsing process in that readers might reanalyze it to reach the more suitable interpretation. But strictly speaking, an ambiguous sentence is not a typical garden path sentence.

2.3 Summary

After observing the missgemented phenomenon of machine, we believe that there is huge distance between computer and human brain. According to Minidxer (2008), *yan jiu sheng ming* can be divided into *yan-jiu-sheng/ming* “graduate student/life” or *yan-jiu/sheng-ming* “study/life”, segmentation of the latter can be more assured and determined apparently if it is human brain, but it is pretty hard for a PC to make it happen. Human brain is as same as PC in the part that it will also present different segmentations while analyzing overlapping and combination ambiguous strings. However, human brain does not missegment the strings but just reports different ways of comprehension.

The study deals with language processing. By definition, language processing refers to the way human beings process speech or writing and understand it as language. Hence, language comprehension of people is included in this field. Instead of studying language processing of machine, the study gets back to real person to understand primary comprehension base of human brain on language and to further find out possible rules for Chinese word segmentation.

3. METHODOLOGY

The study will conduct questionnaire survey. With the assumption that a compound will lead to ambiguity easily, almost all of the sentences in the questionnaire contain compounds. To test readers’ comprehension, two interpretations are provided for interviewees to choose from. 300 students are invited to conduct semantic selection. Based upon the survey results, we will try to figure out the possible criteria interfering while the students are processing the sentences. The Optimality Theory will eventually be adopted to propose the possible ranking of the criteria.

4. RESULTS & DISCUSSION

The result is as follows.

Table 1
xxxxxx

NO.	Sentences	Interpretations	Total (person)	Percentage
1	Wo yong xin suan, zhongyu jiechu le da-an.	I calculate attentively and finally get the answer.	196	65%
		I utilize mental calculation to get the answer.	104	35%
2	Da hua chi deng le yi nian.	Da-hua waited for someone desperately for one year.	111	37%
		Such person is a boy crazy, she has waited for someone for one year.	189	63%
3	Na jia dian you mai su shi ji.	That store sells chicken eating vegetable.	64	21%
		That store sells food made of bean curb but with a taste of chicken.	236	79%
4	Ren hao duo, hao nan guo.	I am upset about too many people.	98	33%
		There are too many people to walk through.	202	67%
5	Xiaoming shi da ren	Xiaoming is a giant person.	7	2%
		Xiaoming is an adult.	293	98%
6	Yong sui zhu dan bijiao fangbian.	It is easier to boil egg with boiling water.	155	52%
		An egg boiled with water is more convenient.	145	48%
7	Ta nashou xiangmu shi la mian.	He is good at making hand made noodles.	30	10%
		He is skilled at cooking Ramen.	270	90%
8	Xiao xin gan	Watch out your liver!	101	34%
		someone important to oneself	199	66%
9	Xuexiao tichang an quan xing xing wei.	The school advocates doing safe things.	26	9%
		The school advocates a precaution against sexual intercourse.	274	91%
10	Ta shi tai ping gong zhu, suoyi hen zibei.	She is Princess Peace in Tang Dynasty, so she feels inferior.	4	1%
		She feels inferior because of her flat breast.	296	99%

Basically, the two interpretations provided for each sentence in the questionnaire contain the reading of either a compound or a phrase. The following displays the possible compounds, phrases and proper nouns in the questionnaire.

Table 2
xxxxx

Compound	**yongxin (attentively),** xinsuan (mental calculation), **huachi (boycrazy),** nanguo (sad), **daren (adult)**, suizhudan (boiled egg), **lamian (Ramen)**, xiaoxin (watch out), **xingan (darling)**, anquanxing (safety), **xingxingwei (sexual intercourse), sushi (food made of bean curd)**
Phrase	chi deng (to wait desperately), **nan guo (hard to get through),** da ren (a giant person), **yong sui (to use water)**, la mian (to make hand-made noodles), **tai ping gongzhu (girls with flat breast)**, su shi (to eat vegetable foods)
Proper noun	Dahua (Ms. Flower), Taiping gongzhu (Princess of Peace)

The morphemes in bold in the table are the choices of most people. According to initial observation, compound reports the most choices, indicating that most people comprehend the strings as compounds.

4.1 Discussion

The questionnaire reports the same result as artificial intelligence analyzer in that there are more than one solution for reading combination ambiguity strings and overlapping ambiguity strings. However, different from machine, human brain did not missegment these strings, they just report different tendencies in comprehension. Basically, the two interpretations of every sentence are reasonable. However, people seem to show preference for one to the other. We hereby introduce descriptive grammar to describe the selection tendency of testers and try to discover the psychological base that masters comprehension as below.

A. Combination Ambiguity String

In a combination ambiguity string, the composed elements could either appear independently as free morphemes or combine together to form another morpheme. Suppose there are two elements A and B. If A and B are regarded as independent words, the phrase is the combination result while meaning is the combination of both. It is a compound with specific meaning if AB is a morpheme. In the following, the five strings could be either a phrase or a compound.

Table 3
xxxxx

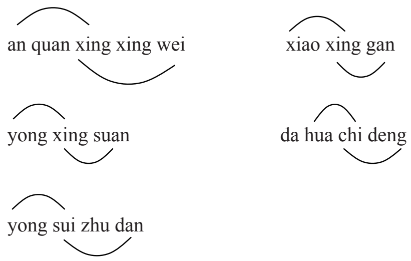
	Phrase	Compound
da ren	giant man	ˇadult
su shi	vegetarian	ˇfood made of bean curd
la mian	make hand made noodles	ˇRamen
nan guo	ˇhard to get through	sad
tai ping gong zhu	ˇfemale with flat breast	Name of person

According to the result of the survey, there are more options of the compound than options of the phrase for the strings *da ren*, *su shi* and *la mian*. *Nan guo* and *tai ping gong zhu* reported the opposite result: more options of the phrase than the compound. We tick in front of those with more options.

B. Overlapping Combination String

Overlapping combination string contains at least three elements, and the parsing of the one(s) in between will lead to ambiguity. Suppose there are three elements A, B and C. B could be combined with A (AB/C) or C (A/BC). Dealing with this type of strings, machine will easily cause missegmentation while human brain will cause ambiguity phenomenon easily. In the questionnaire, the strings *an quan xing xing wei*, *xiao xin gan*, *yong xin suan*, *da hua chi deng* and *yong sui zhu dan* belong to this type.

The string *an quan xing xing wei* can be interpreted as *an-quan-xing*[2] "safety" and *xing-wei* "behavior" or *an-quan* "safe" and *xing-xing-wei* "sexual intercourse". *Xiao xin gan* can mean either "watch out your liver" or "sweetheart", in which the first reading is to make *xiao-xin* as a compound verb and *gan* as a noun and the second reading is to combine *xin* "heart" and *gan* "liver" as a compound noun first and then a prefix *xiao* "little" to signify intimacy. As for the string *yong xin suan*, *xin* could be combined with *yong* to form a compound adverb[3] to modify the verb *suan* "calculate" or with *suan* to form a noun "calculator" as the object of the verb *yong* "use". *Da hua chi deng* can be interpreted in two ways". First, the prefix *da* "big" and the compound *hua chi* "boy crazy" combine to be the subject of entire sentence, and *deng* "wait" is the main verb; the second, *Da-Hua* appears as a proper noun and *chi* modifies the main verb *deng* to mean "waits desperately". *Yong sui zhu dan* seems to have the same verb-object structure as *yong xin suan*: *yong+xin-suan* "use mental calculation" and *yong+sui-zhu-dan* "use boiled egg". Another way of segmenting the string *yong sui zhu dan* is *yong/sui/zhu/dan*. *Yong-sui* (to use water) and *zhu-dan* (to boil eggs) are phrases instead of compound verbs.



The results of questionnaire are as below:

yong-xin (attentively) > xin-suan (mental calculation)
xin-gan (darling) > xiao-xin (watch out)
xing-xing-wei (sexual intercourse) > an-quan-xing (safety)
hua-chi (boy crazy) > chi deng (wait desperately)
yong sui (use water) > sui-zhu-dan (boiled eggs)

The first three pairs are competitions between two compounds while the latter two are competitions between a compound and a phrase. The result of competition between compound and phrase comprised half for each: the compound *hua-chi* "boy crazy" outperforms the phrase *chi-deng* "to wait desperately" whereas the phrase *yong-sui* "use water" outperforms the compound *sui-zhu-dan* "boiled eggs". In our opinions, the compound *sui-zhu-dan* is not chosen by most people probably because of its low frequency of use. Actually, the word *sui-zhu-dan* is not found in Sinica Corpus. The low frequency of appearance then decreases the probability of selection in language processing.

4.2 Rules

Is there a rule existing in abovementioned outcome? There seems no. As for a combination ambiguity string, it seems that people do not show preference for compounds or phrases. In Table 4, people sometimes pick up the reading of compound (e.g., *su-shi*, *da-ren*, *la-mian*) and sometimes that of phrase (e.g.*,nan guo*, *yong sui*, *tai ping gong-zhu*). Even in an overlapping ambiguity string, the element in the middle is not by all means part of a compound or a phrase.

As for a competition between two compounds in an overlapping ambiguity string, does the one surpasses the other report some characteristics? We try to get started from word frequency, assuming that the higher frequency the compounds appear, the easier they are to be chosen. According to Sinica Corpus, the word frequency of *yong-xin* "attentively" is higher than *xin-suan* "mental calculation".

[2] *Anquanxing* basically consists of the compound *anquan* and the suffix *xing*.

[3] According to Zhao, *yong-xin*"is a verb-object compound which is used as an adverb.
a It can be modified by degree adverb. tai yong-xin
b It can take the suffix –de. yong-xin-de xie
c It can be reduplicated. yong-yong-xin-xin-de xie

Word frequency: Yong-xin

No	Rank	Word	Frequency	Percent	Cumulation
3145	3138	yong-xin (VH)	175	0.004	73.336
20076	19413	yong-xin (Na)	18	0.000	90.686
48837	44359	yong-xin (Nv)	5	0.000	96.038

Word frequency: Xin-suan

No	Rank	Word	Frequency	Percent	Cumulation
24337	23198	xin-suan (Na)	14	0.000	92.038
148976	93826	Xin-suan (VA)	1	0.000	99.661

Xing-xing-wei “sexual intercourse” presents higher frequency than *an-quan-xing* “safety”.

Word frequency: An-quan-xing

No	Rank	Word	Frequency	Percent	Cumulation
7934	7855	an-quan-xing (Na)	58	0.001	82.894

Word frequency: Xing-xing-wei

No	Rank	Word	Frequency	Percent	Cumulation
6808	6749	xing-xing-wei (Na)	70	0.001	81.431

Xiao-xin “watch out” appears more frequently then *xin-gan* “darling”.

Word frequency: Xin-gan

No	Rank	Word	Frequency	Percent	Cumulation
39372	36388	xin-gan (Na)	7	0.000	94.957

Word frequency: Xiao-xin

No	Rank	Word	Frequency	Percent	Cumulation
1443	1439	xiao-xin (VK)	430	0.009	64.035
68829	57995	xiao-xin (Nv)	3	0.000	97.478

The former two examples comply with our hypotheses; however, *xiao-xin-gan* reports inconsistent result.

4.3 Optimality Theory

Optimality theory (frequently abbreviated **OT**) was originally proposed by the linguists Alan Prince and Paul Smolensky in 1993, and later expanded by Prince and John J. McCarthy. It is a linguistic model proposing that the observed forms of language arise from the interaction between conflicting constraints. There are three basic components of the theory:

i. GEN takes an input, and generates an infinite number of possible outputs or candidates.

ii. CON provides the criteria, in the form of strictly ordered violable constraints, used to decide between candidates. The higher order the constraint, the more important it is.

iii. EVAL chooses the optimal candidate based on the constraints, and this candidate is the output.

Phonology is the area to which optimality theory was first applied. Although much of the interest in optimality theory has been associated with its use in phonology, the theory is also applicable to other subfields of linguistics (e.g., syntax and semantics).

We try to adopt optimality theory to figure out possible reason for the violation of *xiao xin gan*. Let’s firstly default the word frequency as constraint, the options of each item on below table report high frequency are shown on the left (Sinica Corpus based), while on the right are the options that outperformed in the questionnaire survey. Basically, word frequency only applies to words but not phrases in that it si hard to count the frequency of phrases. As to an overlapping ambiguity string, the frequency of the compound use can be counted, but the frequency of the phrasal use cannot be figured out, so that both parties cannot be compared with each other. In the following, we mark the overlapping ambiguity strings with ※to indicate inapplicability.

Table 4
xxxxx

	High term frequency	Questionnaire result
1	yong-xin	yong-xin
2	hua-chi	hua-chi
3	※	su-shi (food made with bean curd)
4	※	nan guo
5	※	da-ren (adult)
6	※	yong sui
7	※	la-mian (Ramen)
8	xiao-xin	xin-gan
9	xing-xing-wei	xing-xing-wei
10	※	tai-ping gong-zhu

Word frequency could only make a correct prediction for three sentences, while the rest which fail to figure out frequency cannot be adopted. Therefore, we propose another constraint with higher ranking-- “idiomaticity”, indicating that the meaning is specific and cannot be directly told from the composed elements.

Table 5
xxxxx

		Idiomaticity	High word frequency	Questionnaire result
1	☞yongxin	*		yongxin
	xinsuan	*	*	
2	☞huachi			huachi
	chideng	*	*	
3	☞sushi			sushi
	su shi	*	*1	
4	☞nanguo			
	nan guo	*	*	nan guo
5	☞daren			daren
	da ren	*	*	
6	☞yongsui	*		yongsui
	suizhudan	*	*	
7	☞lamian			lamian
	la mian	*	*	
8	xiaoxin	*		
	☞xingan		*	xingan
9	anquanxing	*	*	
	☞xingxingwei	*		xingxingwei
10	☞taiping gongzhu			
	Tai ping gongzhu	*	*	Tai ping gongzhu

We think the compounds *sui-zhu-dan* and *xin-suan* violates the constraint because the meanings of them is easily got from the components; therefore, they are marked with *. We precisely predict the results of eight sentences after the constraint idiomaticity is added on. As for the two sentences with wrong prediction, we believe that the choice of respondents is probably affected by linguistic context. Many scholars have proposed methods to eliminate ambiguity, and linguistic context is part of them. To observe carefully, the fourth and the 10th sentences do offer clear contexts for interpretation.

Ren hao duo, hao nan guo…
Ta shi tai ping gongzhu, **suoyi hen zibei**.

There is direct correlation between "bulky people" and "hard to get through". The Princess Peace in Tang Dynasty does not have to feel inferior than others, so the term "inferiority" will easily associated with small breast.

Linguistic context should be on the top rank of all constraints, and the outcomes of all ambiguous sentences can be precisely predicted after the constraint of linguistic context is added on.

Table 6
XXXXX

		Context	Idiomaticity	High Word Frequency	Questionnaire
1	☞yongxin		*		yongxin
	xinsuan		*	*!	
2	☞huachi				huachi
	chideng		*!	*!	
3	☞sushi				sushi
	su shi		*!	*!	
4	nanguo	*!			
	☞nan guo		*!	*!	nan guo
5	☞daren				daren
	da ren		*!	*!	
6	☞yongsui		*		yongsui
	suizhudan		*	*!	
7	☞lamian				lamian
	la mian		*!	*!	
8	xiaoxin		*!		
	☞xingan			*!	xingan
9	anquanxing		*	*!	
	☞xingxingwei		*		xingxingwei
10	taiping gongzhu	*!			
	☞Tai ping gongzhu		*!	*!	Tai ping gongzhu

CONCLUSIONS

The study conducts questionnaire survey upon ambiguity phenomenon with regard to Chinese compound, it is found that the strings that cause ambiguity are as same as strings missegmented by artificial intelligence, meaning that such type of Chinese string is hard to manage. Artificial intelligence would easily misinterprete the strings because of the missegmentation while analyzing or translating language, consequently, it will eventually lead to lower credibility in language processing.

Currently there are three common ways of word segmentation for PC, and the three constraints proposed by us justly correspond to the three word segmentations of PC. Linguistic context requires larger language information and is mainly about comprehension; idiomaticity is extremely correlated with dictionary, in which those listed in a dictionary are not phrases but normally terms with specific meaning; word frequency is just a key to word segmentation that statistic word segmentation subjects to. PC already adopts the three criteria we propose, but the error rate of segmentation is still high. Why? In our opinions, the key is that the three criteria should be applied simultaneously but not separately. Each of the three segmentation methods adopts one criterion respectively. The processing result would be more convincing if the three criteria are adopted simultaneously with the ranking of importance. There is certain procedure and rule for human brain to follow while processing languages. It will largely enhance the accuracy of language processing for sure if applying the mechanism of human brain's language processing to artificial intelligence.

Language learning is not just familiarity with pronunciation, vocabulary and grammar, meaning communication and comprehension are the ultimate purposes. Therefore, how to correctly express meaning and how to interpret the meaning precisely are the important parts of language learning. Through the study, we understood the important base of general people in understanding ambiguous sentences, which can be provided as a reference for students to deal with ambiguous sentences while conducting language teaching.

REFERENCES

Chen, X. (2010). Cong Yuyixue Jiaodu Tan Hanyu zhong de Qiyi Xianxiang [Talking about ambiguity in mandarin from semantic perspective]. Retrieved from http://www.jiaokedu.com/discourse/hyywx/278505.html

Cui, Z. C. (2008). Lüe Lun Xiandai Hanyu de Qiyi Xianxiang [General study on ambiguity in modern mandarin]. *The Science Education Article Collects*, 33.

Huang, C. N. (1997). Segmentation problems in Chinese processing. *Applied Linguistics, 1*, 72-78.

Li, C. N., & Thompson, S. A. (1981). *Mandarin Chinese: A functional reference grammar*. Taipei: The Crane Published.

Lü, S. X. (1984). Qiyi Leili (Study on Ambiguous Sentences). *Zhongguo Yuwen*, 5.

McCarthy, J. J. (2002). *A thematic guide to optimality theory*. Cambridge: Cambridge University.

Minidxer. (2008). Retrieved from http://blog.minidx.com/2008/01/04/352.html

Prince, A., & Smolensky, P. (2004). *Optimality theory: Constraint interaction in generative grammar*. Blackwell.

Sun, M. S., & Zuo, Z. P. (1998). Overlapping ambiguities in Chinese text. In *Overlapping ambiguities in Chinese text* (pp.323-338).

Wei, Q., Sun, M.S., & Menzel, W. (2008). Statistical properties of overlapping ambiguities in Chinese word segmentation and a strategy for their disambiguation. *TSD '08 Proceedings of the 11th international conference on Text, Speech and Dialogue*.

Xu, H. Z. (2006). *Duoyi yu Qiyi—Taiwan Guanggao Yuyan Shili Fenxi [Polysemy and Ambiguity—Analysis on Advertisement Language in Taiwan]* (Master's thesis). Taipei: National Chengchi University.

Xu, S. Y. (1985). Zai Yiding Yujing zhong Chansheng de Qiyi Xianxiang [Ambiguity in certain context]. *Zhongguo Yuwen*, *5*.

Zhang, L. (2009). Jianlun Hanhu Qiyi Xianxiang [Brief study on ambiguity]. *Journal of Yangtze Normal University*, *25*(5), 133-135.

Zhao, Y. R. (1992). Hanyu zhong de Qiyi Xianxiang [Ambiguity in mandarin]. In *Zhongguo Xiandai Yuyanxue de Kaituo han Fazhan* (pp.249-263). Bejing: Tsinhua University Press.

Zhu, D. X. (1980). *Xiandai Hanyu Yufa Yanjiu* [The research of modern mandarin syntax]. Beijing: Shangwu Yinshu Guan.