中央研究院歷史語言研究所集刊
第六十九本，第一分
出版日期：民國八十七年三月

A Data-driven Approach to the Mental Lexicon: Two Studies on Chinese Corpus Linguistics

Chu-Ren Huang,* Kathleen Ahrens,** and Keh-jiann Chen***

In this paper, we attempt to show i) that corpora offer real instances of language use (production) in a non-controlled environment, ii) that corpora constitute of a large sampling of the real input to linguistic perception, and iii) that corpora extracted from mass media represent the shared linguistic information of the language-speaking community.

Corpus-based studies are studies of linguistic theories based on linguistic objects (instead of on non-linguistic acts like naming, picture pointing, story-telling, or making decisions on yes-no questions.) We use two corpus-based studies to show that they can complement the traditional psychology-oriented studies based on controlled experiments. The two studies shed important light on the psychological reality of the notion of a word in the mental lexicon.

Our first study examines the definition of compounds based on M.I. (mutual information) values extracted from a corpus. We show that this empirically based definition of compounds easily resolves the previous controversies involving intuitive judgements (e.g. Bates et al. 1992 and 1993, and Zhou et al. 1993).

The second study involves the complex cognitive process of *suo1xie3* (abbreviation) and a simple statistical model. We show that while a rule-based model can only capture incomplete aspects of Chinese abbreviation, corpus-based statistical values nicely reflect their status in the mental lexicon.

In conclusion, we argue that corpora reflect shared uses of language and are efficient tools for establishing baseline facts in (psycho-/neuro-)linguistic research.

Keywords: Mental lexicon, Corpus, Word, Mutual information, Abbreviation

---

* Institute of Linguistics, Academia Sinica
** Department of Foreign Languages and Literature, National Taiwan University
*** Institute of Information Science, Academia Sinica

I. Introduction

Sapir (1921) defines language as 'a purely human ... method of communicating ideas, emotions, and desires by means of a system of voluntarily produced symbols.' That is, linguistic objects are objects that carry meaning, and linguistic acts are acts performed to communicate meaning among speakers. Thus, the grammar of a language can be defined as the shared knowledge of the speakers of that language (Huang 1994). As attested by polyglots, speakers of different languages do not differ in the structure of their mind but in the linguistic knowledge that they share with each linguistic community. Adopting the view that linguistic competence is the shared and structured knowledge of the speakers, it follows that linguistic competence can be deduced by 1) extracting the structured knowledge over a vast amount of linguistic objects, as well as by 2) induction from the observation of the mental/ cognitive process.

An example of the latter method is a standard psycho-/neurological study of language, where a group of subjects are tested in a controlled environment, preferably a lab. The tests are carefully designed and controlled, and the parameters monitored are quantifiable attributes. The quantified results include counting of (in)correct answers, reaction time to certain stimulus, as well as interpretations of brain waves in ERP (Event-Related Potential) studies. In other words, these experiments are indirect probes of natural languages restricted by the laboratory environments. As observed in Chafe (1992), the cognitive scientists chose to limit their domain of study for good reason: the basic scientific criterion of verifiability.

In addition to satisfying the standard of verifiability, the experiment-driven approach of cognitive psychologists is also successful in hypotheses testing. In general, there are two approaches to decoding the processes in a black box: to observe the output of the black box, and/or to correlate a black box process to a corresponding measurable process. The quantified parameters mentioned in the preceding paragraph are good examples of the latter. The reactions to a stimulus, such as pointing to a correct picture, or pushing a button, or the surge in a certain brain wave,

are not linguistic processes per se. But by measuring these reactions, hypotheses can be made about the correlated content, duration, and location of the linguistic process itself.

Moreover, given the vastness of natural language (the fact that the observable outputs are infinite) and the earlier difficulty behavioral psychologists had in reaching substantive results by directly quantifying linguistic behaviors, the experiment-driven approach does seem to be the best alternative and has moved the field into promising areas of research in language processing. Naturally, the success of this experiment-driven approach, like any other experiment-driven approach, leads to the question of how the laboratory results can be interpreted in relation to the real world of linguistic acts involving people in non-controlled environments. This question brings us back to the first method of deducing linguistic competence that we mentioned above: extracting the structured knowledge of speakers over a vast amount of linguistic objects, such as a large corpus database. It is the goal of this paper to show that the experiment-driven approach can be complemented by a data-driven approach to approximate the psychological reality of language.

The strongest case for the data-driven approach to linguistic competence can be made regarding the lexicon. The meaning and form of each and every lexical item are as such only because of the shared knowledge of the speaking community. The form *men* refers to 'door' in Mandarin and 'man (plural)' in English. The same concept of a common cooking vessel shared by Chinese cultures are referred to as *kuo* in Mandarin, *diã* in Taiwanese, and *wok* in English (through borrowing from Cantonese). In other words, the relation between a **signifier** and the corresponding **signified** is arbitrarily conventionalized, as succinctly described by Saussure. The shared linguistic knowledge that defines the grammar of a language must include the knowledge of these conventionalized relations. We will assume that this knowledge is the essential part of the mental lexicon. In addition, we will assume that the mental lexicon is shared by speakers of the same language.

In this paper, we will argue that the recent developments in corpus linguistics offer a new and efficient approach that allows cognitive scientists to look into

linguistic processes as they are and as the sum of the society that uses them. The two studies reported here both involve the structure of the mental lexicon. In the first study, we will show that a word is an empirically verifiable unit even in a language which does not conventionalize wordbreaks in its writing system. In the second study, we will show that a complex word-formation process which seems to involve several competing cognitive principles can be reduced to one generalization extracted from the data. In addition, we will show that corpora will offer a more accurate estimation of word probability and degrees of word association, both crucial factors in designing psycho- and neuro-linguistic experiments. First, however, we will introduce our data and the tools of exploration used on this data.

II. Preliminaries: The Academia Sinica Corpus and Theoretical Foundations of Research Tools

II.1 The Academia Sinica Corpus

A corpus can be defined as a collection of standardized electronic texts selected according to a set of design criteria (e.g. Atkins, Clear, and Ostler 1992). The unmarked case of corpus design criteria is to select a collection of data that reflects the behaviors of a language in general, such as the case of the first *bona fide* corpus, the Brown corpus (Kucera and Francis 1967). However, recent studies find that the size of the corpora is an even more fundamental criterion. This is because the amount of data available in a corpus inevitably restricts the possibility of observation. A single observation can hardly suffice as a foundation for the description of the behavior of an item. In order to increase the population of valid observations, a corpus must at least contain a certain number of minimal instances of the item. Thus, the recent trend in the field of corpus linguistics is to favor corpus size over balanced selection (e.g. Church and Mercer 1993), even though the debate between balance and size is far from settled. The Academia Sinica corpus

that we use in this study is a good example of unbalanced corpus with sufficient size to reflect linguistic generalizations.[1]

The Academia Sinica Modern Mandarin Corpus is developed as part of the on-going research on computational and theoretical Chinese linguistics at the Chinese Knowledge Information Processing Group (Huang and Chen 1990, and Chen and CKIP 1991). It consists of 40 million characters of written texts. The texts are contemporary written texts and are mainly extracted from news journals. Of the texts, 20 million characters are connected on-line with our corpus research tools. The remaining texts are on disks ready to be ported for expansion and/or testing. A 20 millions character corpus is equivalent to roughly 14 million words, the size of the complete COBUILD English corpus. This is also the size of other (mainly English) corpora that have yielded reliable linguistic generalizations (e.g. Sinclair 1987, and Church and Mercer 1993).

The rough breakdown of the texts in the Academia Sinica Modern Chinese corpus is given below.

i. 8 million characters of Liberal Times, a daily newspaper.
ii. 30 million characters from China Times.
iii. 2 million characters from other newspapers, included United Daily and Children's Daily.
iv. 75 thousand characters of standard grade school Mandarin textbooks. This represents the basic vocabulary and sentence patterns that all Mandarin speakers in Taiwan share.
v. 155 thousand characters from Dr. Sun Yat-Sen's transcribed speech. This serves the dual purpose of both representing spoken language text and supplying some of the basic political vocabulary used in Taiwan.
vi. Other miscellaneous texts include subtitles of a TV talk show, technical reports, theses, and magazine articles.

---

[1] Sinica Corpus, a smaller balanced corpus fully tagged with grammatical categories has just been completed at Academia Sinica. This fully tagged balanced corpus contains two million words. Please see Huang et al. (1995) for more detailed discussion.

It should be clear that, though unbalanced, this corpus does present a fairly representative sampling of the reading perception of an average Chinese speaker in Taiwan. Newspapers are the mainstay of our data source for the obvious reason that the newspaper texts are on-line. They also happen to be a rich source of different types of texts, because they contain many varieties of writing, including spoken conversation (interviews), commentaries, translations (foreign dispatches), and all genres of literary styles (Chinese newspapers are different from Western ones in that they have daily literature supplements, and are one of the most important venues for literary publications). We are continually acquiring new texts from these journals and other resources.

II.2 Corpus Research Tools

What makes corpus different from the raw, unaccessible, and infinite amount of linguistic data we encounter everyday is that automatic searches can be performed and generalizations can be extracted with computational tools. The two basic tools that are shared by many corpus-based studies are KWIC and collocation search programs. KWIC provides an aid for human experts to locate and extract relevant data in order to induce generalizations, while collocation will extract generalizations in terms of frequency distribution and mutual information for linguists to interpret.

The KWIC (Key-Word-In-Context) search program is in essence an automatic concordancing program. When given a key word (or character sequence), KWIC automatically locates all the contexts containing the specified key in the corpus. It also does the sorting automatically and presents the data according to the condition specified by the linguist. Our current program allows linguists to specify the size of both right-and left-hand side contexts shown as the search result. It also allows linguists to choose from different sub-corpora to control both the size and the domain for the search. The currently available corpora sizes range from 1 million to 20 million characters. Since the left or right context of the key is sorted, generalizations over the uses of that key can be easily detected and hypotheses can be formed accordingly. When tagged, the context can also be sorted according to

the tags, such as grammatical categories. Boolean constraints involving another element, such as co-occurrence to the left of the primary key, can also be provided. All the above research functions are implemented or being implemented on the Academia Sinica corpus.

The collocation program quantifies the correlations between the key and its context based on the KWIC search result. The crucial role a collocational program plays in a corpus underlines both the recent theoretical emphasis on the lexicon as well as the long-held lexicographical motto that a word can be described in terms of the company it keeps (Sinclair 1991). Whether two linguistic elements are collocates or not is calculated in terms of the statistical values of mutual information instead of using frequency of occurrence. The *a priori* counting of frequency is not an effective indicator of collocation because of the disparity among the probability of different words. In our corpus, *de* (3.5%, see CKIP 1993b) is 1,700 times more likely to occur than an average word (0.002%) and would be mistakenly taken as a correlating word in almost all contexts if frequency is adopted as the criterion.

Mutual information, however, controls for this by calculating the probability of the pair (x,y) co-occurring relative to the production of their individual probabilities in the corpus. It is defined as: $I(x,y) = \mathrm{Log}_2 P(x,y) / P(x) \cdot P(y)$, where **I** stands for **mutual information**, and **P** stands for **probability** (Church and Hanks(1990)). The collocation of **x** and **y** are significant only when their co-occurrences are more likely than chance. Thus collocates are defined as the pairs whose mutual information is much larger than 0. In our corpus, mutual information can only be calculated between a key and the characters occurring in the specified window (i.e. the left and/or right context) at present. Word-based mutual information will be implemented shortly. The value of mutual information is approximated according to the following equation, where **N** is the size of the corpus and **m** is the size of the selected window:

$$
\begin{aligned}
I(x, y) &= \mathrm{Log}_2 P(x,y) / P(x) \cdot P(y) \\
&\approx \mathrm{Log}_2 \frac{f(x,y) / m \cdot N}{f(x)/N \cdot f(y)/N} \\
&= \mathrm{Log}_2 f(x,y) \cdot N / m \cdot f(x) \cdot f(y)
\end{aligned}
$$

Using the above equation, we are able to select possible associative characters with a keyword. Our program allows the collocating characters to be ranked both in terms of mutual information values and frequency of occurrence. In either case, linguists have the option to set a threshold to filter out non-significant collocates with either a low mutual information value or a small number of occurrences. Our experience shows that it will be most efficient to use the number of occurrences as the threshold because the accidental presence of a rare character usually entails an artificially high mutual information ranking. Once this is done mutual information is a much more dependable, and theoretically significant, indicator for evaluating collocates.

III. Using Corpus to Attest Wordhood

III.1 Introduction

One important feature of Chinese texts is that they are character-based, not word-based. Each Chinese character stands for one phonological syllable and in most cases represents a morpheme. The fact that Chinese writing does not mark word boundaries poses the unique question of word segmentation in Chinese computational linguistics (Sproat and Shih 1990, and Chen and Liu 1992). Since words are the linguistically significant basic elements that are entered in the lexicon and manipulated by grammar rules, no language processing can be done unless words are identified. Furthermore, since the concept of a word is not conventionalized in writing, potential controversies can arise in linguistic studies involving the concept of words. For instance, the neurolinguistic work of Bates et al. (1991) was challenged not on basis of theoretical ramifications but rather on the basis of what is perceived to be a 'word' (Zhou et al. 1992). We recognize that wordhood is a primary linguistic construct and will try to propose a verifiable method to identify words. The primacy of the concept of word can be more firmly established if its existence can be empirically supported in a language that does not mark it conventionally in texts.

Chinese words may be made up of one or more characters, with more than 65% of lexical entries being made up of two characters and constituting 42% of the total words (Chen et al. 1993).[2] Since words are not marked by boundaries in Chinese, there are many instances where it is difficult to decide whether two characters in Chinese are one word or two. For example, compare *shou-biao* 'hand-clock' (watch) and *jia-biao* 'fake-clock' (fake-watch). Most native speakers would feel that *shou-biao* is one 'word' while *jia-biao* is made up of two 'words'. But how do native speakers know this? Or more importantly for our purposes: how can linguists capture this native speaker intuition? We will show that multiple character words in Chinese can be identified by examining their distribution of characters in a large corpus, which is the collection of attested instances of language production as well as the shared targets of language comprehension.

III.2 V-N Compounds vs. V-N Phrases: Processing Implications

The fact that a word is not a conventionalized psychological concept in Chinese has implications for examining the way language is organized, accessed and processed in the brain. For example, in instances where one would like to examine sublexical processes, it is difficult to do so if one cannot first ascertain what constitutes a lexical item. Such a dilemma can be found in a recent neurolinguistic experiment on noun-verb dissociation in Chinese aphasics (Bates et al. 1991).

The experiment tested the breakdown in the production of nouns and verbs in Broca's and Wernicke's aphasics and found that a double dissociation exists between object and action naming for these two subject groups. A morphological account of this phenomenon has been ruled out because scholars had postulated (on the basis of neurolinguistic studies in European languages) that it is the morphological complexity of the verb that hampers Broca's production, and Chinese lacks a morphological complexity difference between these two word types.

---

[2] On the other hand, single character words have less entries but are more dominant in terms of frequency (nearly 54% of total occurrences). See Chen et al. (1993) for more details.

More importantly for current discussion, Bates et al. (1991) also found that this noun-verb dissociation extends to the sublexical level. That is, Broca's aphasics have difficulties producing the verb portion of a V-N compound, while Wernicke's aphasics have difficulty producing the noun portion. This finding, if it holds, argues against both a syntactic and lexical account of the breakdown, and instead supports a semantic conceptual one. However, questions have been raised concerning whether or not the V-N compounds used in the experiment by Bates et al. were really compounds (e.g. 'words') or whether they were in fact 'phrases' (Zhou et al. 1992). If they are phrases, there is no reason to extend the noun-verb dissociation to the sublexical level. However, the syntactic tests to determine the 'wordhood' of a V-N item do not give clear-cut results (Zhou et al. 1992, Bates et al. 1992). The controversy clearly is attributable to the discrepancy of the introspective judgements of different native speakers.

The fact that a Chinese corpus is character-based makes it useful regarding the study being discussed. First, since all sub-lexical units are represented, the distribution of sub-lexical items in actual use in this language can be observed. Second, the corpus allows us to study the distributional correlations between characters, and thus to determine the wordhood of the character pairs. Words are linguistically defined as a unit in terms of distributional properties. That is, words occur as a holistic unit in languages and the constituents of words cannot be scrambled. Notice that the issue here involves compounds, which by definition are words made up from juxtaposing other words. Hence the fact that the constituent characters can occur alone in isolation is given, and not counterexamples to the wordhood of the compounds. What we need to find out is whether each character binds to one another strongly enough to be considered a unit.

Our study takes the 25 V-N sequences used in Bates et al. (1991) as the population.[3] Observation of the distribution of these pairs is made using the collocation program on our corpus (Huang et al. 1993). The collocation program is

[3] The original study lists 27 V-N compounds. But two V-N compounds are found to be repeated entries and thus eliminated.

designed to show the number of times a character occurs within the designated context of the 'key' character. This information is then used to calculate the mutual information value of each character paired with the key character/word. Mutual Information is widely used in corpus linguistics to measure the degree of association between linguistics units (Church and Hank 1990). It has also been used in Chinese computational linguitstics for the task of word segmentation (Sproat and Shih 1990). The MI value of a pair of linguistic elements is calculated based on the frequency of their co-occurrences and their respective probability in the language, as introduced in the last section. A high MI (mutual information) value indicates that two units are highly associative.

In this experiment, we first take the first character of the purported compounds as the keys and search for all the occurrences of each of them in our corpus. We then collocate each key character with the first character following it. The purpose is to calculate the MI value of the target second character (N) to see if it is highly associative with the key character in this position. The assumption, like that of Sproat and Shih (1990), is that two highly associative immediate neighbors are likely to be considered a unit and thus a word. Next, we take the second character (N) as the key and repeat the same process, except that this time collocation is tested for the position immediately to the left of the key. In other words, we are testing the associativity of two concatenating units from both directions. Even though the MI values are identical by definition, bi-directional tests like this allows us to examine the rank of associativity in terms of both constituents. In other words, the collocation results can be further supported if another pair with a lower MI value is known to be a word. Results of the two experiments are summarized in Table I.

#	Chinese VN	English gloss	Translation	MI>2	MI>2 Freq<10	MI<2
1	*xie-zi*	write-character	to write	X		
2	*song-hua*	send-flower	to give flower	X		
3	*diao-yu*	hook-fish	to fish	X		
4	*hua-hua*	paint-painting	to draw picture	X		
5	*cui-kou-shao*	blow-mouth-whistle	to whistle		X	
6	*la-che*	pull-cart	to pull cart			X
7	*xia-lou-ti*	descend-floor-stair	to go downstairs		X	
8	*shu-tou*	comb-head	to comb hair		X	
9	*chi-fan*	eat-rice	to eat (rice)	X		
10	*shang-lou-ti*	ascend-floor-stair	to go upstairs		X	
11	*cui-la-zhu*	blow-wax-candle	to blow candle			X
12	*ni-shui*	drown-water	to drown	X		
13	*he-shui*	drink-water	to drink	X		
14	*ju-gong*	bow-bow	to bow	X		
15	*tui-che*	push-cart	to push cart			X
16	*pai-shou*	clap-hand	to clap	X		
17	*dian-la-zhu*	point-wax-candle	to light candle		X	
18	*feng-yi-fu*	sew-clothes	to sew			X
19	*hua-tu*	painting-picture	to draw picture	X		
20	*an-men-ling*	push-door-bell	to ring	X		
21	*jian-zhi*	cut-paper	to cut paper	X		
22	*qiao-men*	knock-door	to knock	X		
23	*pao-bu*	run-step	to run	X		
24	*you-yong*	swim-swim	to swim	X		
25	*chang-ge*	sing-song	to sing	X		
Total				16	5	4

Table 1: Mutual Information Value of V-N compounds in Bates et al.'s study

Only four of the 25 claimed compounds are found to have MI values lower than two, the rough threshold of character-based associativity of our corpus. These sequences are found to be non-associative, and unlikely to be words. The other 21 pairs have MI values over two, therefore are highly associative and are likely words. However, five of these V-N sequences (*cui-koushao* 'to whistle', *xia-louti* 'to go downstairs', *shu-tou* 'to comb hair', *shang-louti* 'to go upstairs', and *dian-laju* 'to light a candle') have a low frequency between two and four. We cannot put these sequences in the definitely compounds category since MI values are found to be highly reliable when the frequency of occurrences is higher than ten (Church and Mercer 1993). Even though our intuition collaborate with the high MI values, we will classify them as probable words in accordance with our evidence.[4] The remaining 16 sequences have frequencies ranging from 10 (*an-menling* 'to ring') to 1002 (*diao-yu* 'to fish') and can be unequivocally shown to be words based on the corpus data.

Recall that MI is defined as the ratio of actual co-occurrences of a pair of elements (P(x,y)) as opposed to the binomial probability of their co-occurrence if they are independent of each other (P(x)·P(y)). When **x** and **y** are truly independent, actual frequency is equal (or very close) to the binomial prediction, and the ratio of 1 will produce the MI value of 0 ($\log_2 1 = 0$). Thus, a MI value of 2 means that the pair is 4 times more likely to occur together than normal uncorrelated distribution ($\log_2 4 = 2$). This is significant and comparable to the MI threshold of 2.5 used in identifying word boundaries in Sproat and Shih (1990). The slightly higher threshold of Sproat and Shih (1990) is motivated by the fact that their program is unsupervised when executing and will need higher threshold to ensure precision, which they achieved with a rate of 90%. MI values also tend to be artificially higher when one of the pair is a low frequency element. This is more likely to happen when the corpus is smaller (Sproat and Shih 1990 used a corpus of 2.6 million characters, as compared with our corpus of 20 million characters.) In other words, the threshold value of MI for linguistic

[4] Take *cui-koushao* (blow-whistle, to whistle) for example. There are only five instances of *koushao* in the 20 million character corpus. However, four of them are in the context of *cui-koushao*, and the remaining one *cui-zhe-koushao* 'blow-ASP-whistle'. In other words, all occurrences of this sequences can be considered a word.

associativity is emiprically determined, not theoretically pre-set. Like other statistics-based numbers, the choice of threshold values affect the precision rate of the result. It is promising to find that Sproat and Shih's (1990) study and our two studies converge on the number around 2 for defining the associativity of character constituents of a word in Chinese.[5] The accuracy and robustness of this threshold number is tested and confirmed in a more recent study (Huang 1995).

Another technique that can be used to identify words is to compare the rank of MI values of candidate characters paired with the same key. If the MI value of our target sequence is higher than a known word, then we will be get further support that it is a word. Take the character *ge* in *chang-ge* 'to sing' for example. The three pairs that have higher MI values than it are all words: they are *Yin-ge* 'Yinge, place name in northern Taiwan', *li-ge* 'farewell song' and *wan-ge* 'eulogy', and the words that have MI values lower than *chang-ge* include *shi-ge* 'poetry' and *shen-ge* (hymn). In other words, *chang-ge* clearly falls in the range of associativity of words.

The MI values also correlate with other lexical properties. Let's look at the 16 instances where the frequency is greater than 10 and the MI value greater than 2. The highest MI value of over 17 from this experiment belongs to *jugong* 'to bow'. This verb can be inserted by aspect markers and classifier phrases. However, this property, called 'ionization' in Chao (1968), is a common property of many polysyllabic Chinese verbs and does not necessarily show that a word is not simplex. In fact, our intuition suggests that *jugong* is a simplex word, not a compound, since both characters are bound, neither can occur in any other context, and no individual semantic meaning can be assigned to either of them outside of the one the compound entails. The second highest MI value of over 11 belongs to *youyong* 'to swim'. This is another case of a non-compound because the morpheme *-yong* is bound and can never occur alone except when with *you* (though *you* 'to swim' is free as a verb).[6]

---

[5] Sproat and Shih's (1990) study includes a test on different threshold values. Their result shows that the threshold values of 2 and 2.5 gives the best combined precision and recall rates, and precision drops dramatically below 2 while recall drops dramatically above 2.5.

[6] The fact that these are not compounds is also suggested by the fact that the authors fail to find good word-for-word translations for the second part of the V-N sequence. *Jugong* and *youyong* were translated as 'bow-bow' and 'swim-swim' respectively.

The MI values of the 14 words that can be decomposed into sub-lexical elements range between 2 and 9. Although we do not have a large enough sample to make a definite prediction, it does seem that the MI values of two concatenating linguistic elements correspond to the boundedness of them. For example, we predict 1) that constituents of a word are not free and therefore should have MI values higher than phrases, 2) that words with two bound constituents are naturally more associative than words where only one constituent is bound, and 3) that compounds by definition consist of two free elements and therefore the constituents should be less associative and the MI values should be lower than words with one bound morpheme.

In conclusion, our corpus-based study of the 25 claimed V-N compounds from Bates et al. (1991) presents both clear and quantifiable evidence of wordhood in Chinese and suggests that the concepts of various degrees of lexical relations do exist in the mental lexicon and are reflected in language use. First, we showed that 19 of Bates et al.'s (1991) 25 V-N's are likely compounds, two of them are simplex words, and 4 of them are phrases. Since these 19 instances include nearly 80% of the data, we expect the original results of the Bates experiment to hold up even when the inappropriate data are eliminated. Second, we also showed that the linguistic principle of classifying the concept of words according to the boundedness of their constituents is supported by the distributional properties of the language. The associativity between two (sub)-lexical elements measured in terms of mutual information values seem to reflect their lexical status. The simplex words have higher MI values than words with one bound morpheme, and they in turn have higher MI values than compounds.

III.3 Approximating the concept of a word in the mental lexicon

The concept of a word is primary in a mental lexicon. Even though the fact that wordbreaks are not conventionalized in Chinese writing poses difficulties in psycholinguistic experiment design and controversies in linguistic discussion, we claim that it also offers the unique chance of testing both the verity of the concept as well as how this concept links to psychological reality. In the above study of compounds, we

found that a word can be defined as a sequence of highly correlated concatenating syllables. Furthermore, the theoretical classifications of words according to the boundedness of its components seem to be generally quantifiable in terms of the MI values between the components.

Another phenomenon that allows us to examine the psychological reality of the concept of a word is the identification of unknown words in Chinese. The identification of unknown words in Chinese is a great dilemma for both human and computational processing. Unknown words can only be identified when they are checked and found not to be in the dictionary. However, when words are not demarcated, dictionary look-up is not possible. In other words, in order to identify (and learn) unknown words, it is prerequisite that the concept of a word is present in the mental lexicon in order to identify (possible) words.

The lack of (visual) prompts demarcating words in Chinese allows us to test the concept of word in this language more easily. One simple test conducted at the CKIP project is based on Markov chains. The program examines n-grams (i.e. sequences of n consecutive constituents) in a corpus. All n-grams that occur over a threshold number of times and are not in the dictionary are reported. Examination of the preliminary results shows that this rather simplistic model is very powerful in identifying new words. For example, Chinese proper names, which are notoriously difficult because they are not selected from a finite list like Christian names, or capitalized to indicate proper nounhood, are successfully identified by this method. What this preliminary result suggests is not that people use statistics to identify words, but that words are recognized as such because they recur as a unit in language use and perception. That is, words are perceived as a basic linguistic unit and as a unit in the lexicon not because they cannot be further broken down, but because the distributional property of any sub-lexical unit shows clustering at this level. The corpus gives us a database representing a sizable linguistic production which is commonly perceived and comprehended by a good number of speakers, while the statistical tools help us to quantify and identity this clustering property.

IV. *Suoxie:* Abbreviation as a Lexical Rule

The mental lexicon must contain information regarding the lexical process that can generate new words since the vocabulary of any language is an open set that can be expanded. The assumption that such processes are part of the lexicon also has the advantage of eliminating redundancies in lexical representation.[7] A specific rule in Chinese that derives new lexical items from longer corresponding forms is called *suoxie* 'abbreviation'.

The uniquely Chinese *suoxie* is a productive lexical process that any native speaker acquires but so far cannot be captured in terms of theoretical linguistic concepts. *Suoxie* seems to play a role similar to creating acronyms but at the same time differs from it in substantial ways. Acronym formation in Western languages are governed by some variations of the rule of taking the first segment of each word sequentially from a phrase to form a new word. *Suoxie*, on the other hand, forms a new word by taking one character (i.e. syllable) from each word in a compound/phrase, but not necessarily the first character in each word. For instance, the Institute of History and Philology at Academia Sinica is called *Lishi Yuyan Yanjiusuo* '(lit.) History Language Institute.' The abbreviation is *Shiyushuo*, formed by taking the second, first, and third syllables respectively from the three constituents.

What is notable is not the seeming complexity of this process but that native speakers in general know how to form new (shorter) words by *suoxie*. Their intuition on possible and impossible *suoxie* words are similar to their judgments on possible and impossible words. Since *suoxie* is one of the most productive process of neologisms, we will assume that it is part of the mental lexicon and try to account for it.[8] In what follows, we will discuss several principles proposed and discussed in earlier literature and show that they only offer partial explanations of specific instances. We will then show that a unified account can be given based on generalizations extracted

---

[7] This is the basic premise of many current linguistic theories, such as LFG (Bresnan 1982).

[8] In the CKIP study to identify unknown words described above, we found that *suoxie* is one of the two most productive source of new words, the other one being compounding. Derivational rules are actually not very productive according to our preliminary results.

automatically from the corpus. Preliminary results of this on-going study were reported earlier by Hannan et al. (1993).

First, the simple rule of taking either the leftmost constituent or the head constituent have often been mentioned. A combination of either rule seems to account for the majority of the cases even though it is not clear at all which rule to apply on each individual case. For instance, *Meilijian Hezhongguo* 'America United-States', the United States of America', is abbreviated as *mei-guo* 'U.S.A'. The formation process takes the leftmost character *mei* from the head-less proper name and the head *guo* 'country' from the second word. Nevertheless, it is also true that there are a significant number of abbreviated words that cannot be predicted by either rule. For instance, *Zhonghua hangkong (gongsi) 'China Airlines (company)'* is abbreviated as *Huahang*, where the selected second character from the first word is neither the initial character nor the head of the noun phrase.

Second, one could also postulate that the *suoxie* process selects the character/morpheme with the more restrictive meaning as the following example shows. There are two common terms referring to China: *Zhongguo* '(lit.) the Middle Country' and *Zhonghua* '(lit.) Middle China'. It is interesting to observe that *Zhongguo* is usually abbreviated to *zhong* (as in *zhongyou* for *zhongguo shiyou* 'China Petroleum'), while *Zhonghua* is abbreviated to *hua*, as in *Huahang* 'China Airlines'. This fact is even more intriguing considering the fact that they both share the first morpheme *zhong* 'middle'. This fact precludes the possibility that a semantic meaning pre-selects a target. A plausible explanation for this pair is that the morpheme with the more restrictive meaning is selected in each case. *Zhong* 'middle' modifies *guo* 'country' and is therefore more restrictive, while the proper name *hua* 'China' is in turn more restrictive than the directional term of 'middle'. However, this explanation has the following drawbacks: first, it can only apply to a small number of *suoxie* cases. Second, it conflicts with the first hypothesis that it is a head that is selected, and third, it cannot account for the fact that the same word can be abbreviated differently in different contexts, as we will demonstrate below.

A third possible explanation for the non-occurrence of certain *suoxie* words is that the result of an abbreviation cannot be identical to an existing word.[9] That is, like many other morpho-lexical rules, *suoxie* is constrained by suppletion. Take the place name *Gaoxiong* 'Kaohsiung' for example: it is often abbreviated as *gao*, but when the term *Gaoxiong Zhongxue* 'Kaohsiung High School' is abbreviated, the new word is *Xiongzhong*. This can be explained by the fact that the alternative *gaozhong* is an existing lexical item meaning 'high school'. The same explanation can be said for abbreviating *Gaoxiong Gaoshang* 'Kaohsiung Business High School' to *Xiongshang*. This is because *Gaoshang* 'business high school' is an existing entry.

However, such an explanation is at best incomplete because it offers no account how lexical entries pre-empt each other. That is, it cannot predict which word gets the priority when two share the same character. Futhermore, there are cases of possible lexical preemption where more than two words share the same character (the candidate pre-emptor). What often happens is that *suoxie* will never pick the shared character. Hence, we have a seeming preemption effect without the pre-emptor ever actually occurring. This fact obscures the picture again, since a suppletion account presupposes the existence of certain lexical entries. To sum up, we have shown that none of the (partially) motivated rules are able to account for the range of facts concerning *suoxie*. Instead of adding other competing/complementing rules that work as partial accounts, we will propose a radically different unifying account.

In our study on this complex set of facts, we decided to work on a complete subset of words where the *suoxie* process can be reduced to one binary choice. We choose a set where the process is reduced to determining which of the two syllables from a word can be concatenated with a known second part of the *suoxie*. The domain we investigate is county names in Taiwan. There are all together 16 counties in Taiwan. They all have di-syllabic names and all are unambiguously reduced to one syllable when abbreviated. Since *xian* (county) is mono-syllabic, there is no decision making involved in the second part of the abbreviation. Our

[9] An interesting observation is that this is exactly opposite to the formation of acronyms. Acronyms with forms identical to existing lexical entries are favored. Good examples are EMILY('s List), MADD, SADD, SALT, WASP, and WHO.

hypothesis is that the *suoxie* word formation process is controlled by the collocational properties of the possible *souxie* targets.

Our study uses *xian* as the key character and includes all occurrences of *xian* (county). We choose the collocational window of 5 to the left of the key, the standard window size for checking immediate association as suggested in Church and Hank (1990). Please take note that the actual occurrences of the *suoxie* of each county are excluded and the MI value manually adjusted accordingly. The positions of the occurrences are not weighted. The result is show in Table II.

County Name		Abbreviation	First Character	M. I.	Second Character	M. I.	
Taipei	台北	BeiXian	台 tai	2.64	***北 bei***	0.81	*
Taoyan	桃園	TaoXian	***桃 tao***	4.39	園 yuan	3.32	V
Hsinchu	新竹	ZhuXian	新 xin	0.67	***竹 zhu***	0.66	*
Yilan	宜蘭	YiXian	***宜 yi***	3.11	蘭 lan	3.86	*
Miaoli	苗栗	MiaoXian	***苗 miao***	4.40	栗 li	5.37	V
Taichong	台中	ZhongXian	台 tai	2.64	***中 zhong***	0.19	*
Changhua	彰化	ChangXian	***彰 chang***	4.48	化 hua	2.13	V
Nantao	南投	TouXian	南 nan	0.58	***投 tou***	0.29	*
Hualian	花蓮	HuaXian	***花 hua***	0.95	蓮 lien	4.08	*
Yunlin	雲林	YunXian	***雲 yun***	3.78	林 lin	2.13	V
Jiayi	嘉義	JiaXian	***嘉 jia***	3.49	義 yi	2.43	V
Tainan	台南	NanXian	台 tai	2.64	***南 nan***	0.58	*
Kaohsiong	高雄	GaoXian	***高 gao***	1.33	雄 xiong	3.21	*
Taidong	台東	DongXian	台 tai	2.64	***東 dong***	1.29	*
Pingdong	屏東	PingXian	***屏 ping***	1.00	東 dong	1.29	*
Penghu	澎湖	PengXian	***澎 peng***	3.33	湖 hu	1.92	V

Table II: Character-based Mutual Information w.r.t. to Xian (County)

* : Character with lower M. I. is picked to form a new word.

V : Both characters have M. I. higher than 2, therefore are not desirable targets for new words. Default rule applies to pick the first character.

As shown in Table II, ten of the words are abbreviated to the first syllable while six of them are abbreviated to the second one. Of the 32 tokens occurring in the 16 names, 3 types occur in more than one word. *Tai* occurs in four county names but is never used in *suoxie*. *Nan* and *dong* each occur in two names and is the *suoxie* target for one of the two words respectively.

Our initial hypothesis regarding *suoxie* is that it selects the syllable that is less associative with the other parts of the source word/phrase. This is equivalent to formalizing the fact that contexts play a central role in the *souxie* process, as attested by *souxie* words like *huahang, zhongyou*, and *xiongzhong* discussed above. Taking the *xian* example again, our hypothesis is that when determining the *suoxie* of county names, a speaker will only take into account the distribution of the two candidate characters as related to *xian* 'county'. The correlation (and/or frequency) of these candidates in a non-*xian* context does not affect the perception of the possible target *suoxie* form and is therefore irrelevant. This intuition, based on a loose interpretation of Shannon's Information Theory (Shannon 1948 and Pierce 1980), is that the less associative pair carries higher information load and should be the favored target. In addition, when associativity is measured in terms of mutual information, low associativity can also predict the suppletion effect (since constituents of words are highly associative) and the low frequency effect.

One factor that mutual information cannot control, however, is the occasional structural effect of choosing the first syllable. We can see that a straightforward associativity account of choosing the syllable with the lower MI value fares only slightly better than the structural account of choosing the first syllable (11 out of 16 instead of 10 out of 16). However, when we look at the pairs which are not predicted by lower MI value, we found that both syllables have significant MI values with *xian*. Their MI values are all around or above 2.[10] As mentioned above, the MI value 2 indicates associativity for Mandarin character pairs. In terms of the theory of

[10] The threshold number of 2 is arrived at based on empirical evidence and does not entail any theoretical consequence. This is why we consider the MI value of 1.92 for *hu* (in *Penghu*) to be associative. This position will have to be tested with other corpora to see if it does constantly show high associativity with *xian* as we assume.

communication, we may say that neither candidate carries enough information load to automatically qualify them as the *suoxie* target. In this case, the empirical rule of taking the leftmost constituent kicks in. This is because almost all *suoxie* involves nominals, and nominals in Chinese are predominantly of the Modifier-Head type.

We therefore revise our hypothesis to the following: the *suoxie* target is the least non-associative constituent of a word, or the first syllable when all constituents are associative. This is to make the first syllable rule as the default when the criterion of retaining the most information does not provide a viable candidate. But take note that we account for the first syllable rule not in structural terms but in terms of the distributional fact that constituents in this position are usually less associative, and therefore semantically more restrictive. This revised principle correctly predicts all cases of *suoxie* involving county names. The italicized characters in Table II represent the *souxie* targets, and the checked instances are where the default rule applies.

V. Estimating Lexical Probability

In experiment-oriented studies of language and cognition, it is very important that the probes are controlled for factors that may affect the speed of the linguistic act or the corresponding measurable act being monitored. For instance, it is necessary to control the complexity of the linguistic object when the stimulus is presented in the written form. The standard control for word complexity is length in terms of letters in English. Another factor that is known to affect linguistic processing is the probability of lexical items. The probability can be characterized in two ways, 1) the overall probability and 2) the probability in a certain context. In general, the more probable items are more accessible and will require less processing time. The effect of lexical probability on lexical access is well-studied (e.g. Seidenberg 1989). And the priming effect between associative words (such as *bread* priming *butter*) is also well-known.

The issue here is how to estimate these probability values. It is obvious that we will never be able to identify the real probability since the universe of linguistic objects is not only infinite but also ever increasing. The standard way to estimate

probability is of course by sampling. In terms of lexical probability, a corpus is studied and the probability (**P**) is calculated based on the frequency of occurrence (**f**) and the corpus size (**N**). That is,

$P \approx f/N$.

However, for this estimation to be reasonably accurate, corpus size is critical. We have observed that a single observation of an object can hardly be considered an evidence for its behavior. In other words, if an item occurs in a corpus once, it is more likely that it occurs by chance rather than that its probability is 1/N. However, words in a lexicon are distributed such that the probability drops (rather steeply) in reverse proportion to its rank (Zipf's Law). In CKIP (1993b), which lists frequencies of more than 45 thousand words from a subset of our corpus, the probability of the 9,000th word is estimated to be 0.001%. This estimation should be reasonably accurate since the word occur 95 times in the corpus. Following the empirical result that 10 occurrences of any item offers reliable probability counts (Church and Mercer 1993), this number can be interpreted to mean that any Chinese corpus will have to have the size of more than 1 million words to give reliable probability estimation for the first 9,000 words. In fact, the first 9,000 words represent the most commonly used words since they entail 92% of our total corpus. Based on our estimation of the average wordlength of a little over 1.4 for Chinese (Chen et al. 1993), this translates to a corpus size of more than 1.4 million characters needed to calculate lexical probability effects. We are not aware of any previous Chinese corpus that has the size of 1.4 million words.

On the other hand, lexical probability is not the most interesting linguistic information one can extract from a corpus. Linguists are more interested in the interactions of lexical items. In other words, the correlation of lexical items will offer more linguistically relevant information. Ken Church (p.c.) suggests that MI information is only reliable when more than a few hundred instances of an individual keyword is found in the entire corpus (by KWIC search), and our previous work supports this estimation. Without taking into account the probability of the correlating element, this fact alone requires a corpus larger than 10 million words to offer reliable generalizations regarding binary correlations involving two common lexical items.

The two rough estimations above give us hint of why corpus research lay dormant for nearly 20 years after the Brown Corpus. The Brown Corpus offered reasonable estimates for the probability of the core vocabulary of English but was not able to offer many more generalizations due to the limitations of its size. There was simply not enough technology or resources to deal with corpora with the required size. The recent advancement in computational technologies allowed processing of larger corpora and thus opened up the possibility of new research directions.

While lexical properties up to binary relationship do not constitute the full range of linguistic facts, we think it is important to observe that these cover the most important lexical relations. In other words, modern corpora are now able to offer direct and verifiable observation of lexical properties of languages.

VI. Conclusion

There are three important features of corpora that make them irreplaceable in psycho- and neuro-linguistic studies. First, corpora offer real instances of language use (production) in a non-controlled environment. Second, corpora constitute large sampling of the real input to linguistic perception. Third, corpora extracted from mass media represent the shared linguistic information of the language-speaking community. We have shown that corpus-based studies produce reliable and verifiable generalizations regarding the mental lexicon, as evidenced by our studies on V-N compounds and the *suoxie* process. It is our hope that results of this data-driven approach based on corpora will complement the experiment-oriented approached to shed more light on the nature of the psycho- and neuro-logical basis of language.

Acknowledgements

We would like to thank all colleagues at CKIP, especially Feng-yi Chen, Marie Hannan, Wei-Mei Hong, and Yun-yan Yang, for their help and comments. Studies reported here benefited directly from their help. An earlier version of this paper was presented at the International Conference on the Biological Basis of Language. We

would like to thank Ovid Tzeng and Daisy Hung for organizing and providing this intellectually stimulating forum for academic exchanges. We would like to thank participants of the conference, including Al Liberman and Elizabeth Bates, for their helpful comments. The studies reported here are partially supported by a grant from the Chiang Ching-kuo Foundation of International Scholarly Exchanged and a grant from the National Science Council of the R.O.C.

(Accepted for publication 15 November 1997)

Bibliography

Atkins, Sue, Jeremy Clear, and Nicholas Ostler.

1992 Corpus Design Criteria. *Literary and Linguistic Computing* 7.1: 1-16.

Bates, Elizabeth, Sylvia Chen, Ovid Tzeng, Ping Li, and Meiti Opie.

1991 The Noun-Verb Problem in Chinese Aphasia. *Brain and Language* 41: 203-233.

Bates, Elizabeth, Sylvia Chen, Ping Li, Meiti Opie, and Ovid Tzeng.

1993 Where is the Boundary between Compounds and Phrases in Chinese? *Brain and Language* 45: 94-107.

Bresnan, Joan.

1982 *The Mental Representation of Grammatical Relations*. Cambridge: MIT Press.

Chafe, Wallace.

1992 The Importance of Corpus Linguistics to Understanding the Nature of Language. In Svartvik (1992). 79-97.

Chao, Yuen Ren.

1968 *A Spoken Grammar of Chinese.* Berkeley: UC Berkeley Press.

Chen, Ching-yu, Shu-fen Tseng, Chu-Ren Huang, and Keh-jiann Chen.

1993 Some Distributional Properties of Mandarin Chinese——A Study Based on the Academia Sinica Corpus. *Proceedings of the First Pacific Asia Conference on Formal and Computational Linguistics*. 81-95. Taipei.

Chen, Keh-jiann, and CKIP.

1991 The Chinese Knowledge and Information Project and Chinese Electronic Dictionary [in Chinese]. *Paper presented at the Fourth Joint Chinese-Japan Symposium on Information Technology*. Taipei.

Chen, Keh-jiann, CKIP, and Chu-Ren Huang.

1990 Information-based Case Grammar. *Proceedings of the 1990 International Conference on Computational Linguistics (COLING-90)* 2: 54-59.

Chen, Keh-jiann, CKIP, and Shin-Huan Liu.

1992 Word Identification for Mandarin Chinese Sentences. *The Proceedings of the 1992 International Conference on Computational Linguistics (COLING-92)*. 101-105. Nantes, France.

Chinese Knowledge Information Processing Group.

1993a *Corpus-Based Frequency Count of Characters in Journal Chinese* [In Chinese]. *CKIP Technical Report no. 93-01*. Taipei: Academia Sinica.

1993b *Corpus-Based Frequency Count of Words in Journal Chinese* [In Chinese]. *CKIP Technical Report no. 93-02*. Taipei: Academia Sinica.

1993c *The CKIP Categorical Classification of Mandarin Chinese* [In Chinese]. *CKIP Technical Report no. 93-05*. Taipei: Academia Sinica.

Church, Kenneth.

1988 A Stochastic Parser Program and Noun Phrase Parser for Unrestricted Text. *Second Conference on Applied Natural Language Processing*. Austin, Texas.

Church, Kenneth and Patrick Hanks.

1990 Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16.1: 22-29.

Church, Kenneth and Robert L. Mercer.

1993 Introduction to the Special Issue on Computational Linguistics Using Large Corpora. *Computational Linguistics* 19.1: 1-24.

Fillmore, Charles.

1992 "Corpus Linguistics" or "Computer-Aided Armchair Linguistics". In Svartvik (1992). 35-60.

Hannan, Marie-Louise, and Kathleen Ahrens.

1993 Corpus Linguistics on Firm Ground. *Newsletter of the International Association of Chinese Linguistics*. 2.1.

Hannan, Marie-Louise, Kathleen Ahrens, Feng-yi Chen, and Keh-jiann Chen.

1993 A Corpus-based Study of Abbreviation. *Presented at the Second International Conference on Chinese Linguistics (ICCL-II)*. Paris. June 23-35.

Huang, Chu-Ren.

1994 Corpus-based Studies of Mandarin Chinese: Foundational Issues and Preliminary Results. In Matthew Y. Chen and Ovid J-L. Tzeng Eds. In Honor of William S.-Y. Wang. *Interdisciplinary Studies on Language and Language Change*. 165-186. Taipei: Pyramid.

1995 The Morpho-lexical Meaning of Mutual Information: A Corpus-based Approach Towards a Definition of Mandarin Words. *Presented at the 1995 Linguistics Society of America Annual Meeting*. New Orleans, January 5-8.

Huang, Chu-Ren and Keh-jiann Chen.

1992 A Chinese Corpus for Linguistic Research. *In the Proceedings of the 1992 International Conference on Computational Linguistics (COLING-92)*. 1214-1217.

Huang, Chu-Ren, Keh-jiann Chen, Paul M. Thompson, and Pei-Chuan Wei.

1993 Chinese Linguistic Computing: Modern and Classical Chinese Corpora at Academia Sinica. *Proceedings of the Fifth China-Japan Symposium on Information Technology*. Taipei: National Science Council.

Huang, Chu-Ren, Keh-jiann Chen, Li-ping Chang, and Hui-li Hsu.

1995 An Introduction to Academia Sinica Balanced Corpus [In Chinese]. *Proceedings of ROCLING VIII*. 81-99.

Karlgren, Hans.

1990 [Ed.] *The Proceedings of the 1990 International Conference on Computational Linguistics (COLING-90)*. Helsinki, Finland.

Kučera, H. and W. N. Francis.

1967 *Computational Analysis of Present-Day American English*. Providence: Brown University Press.

Pierce, John R.

1980 *An Introduction to Information Theory: Symbols, Signals and Noise*. Second, Revised Edition. New York: Dover.

Sapir, Edward.

1921 *Language: An Introduction to the Study of Speech*. New York: Harcourt Brace.

Seidenberg, Mark S.

1989 Visual Word Recognition and Pronunciation: A Computational Model and Its Implications. In William Marslen-Wilson. Ed. *Lexical Representation and Process*. 25-74. Cambridge: MIT Press.

Shannon, Claude.

1948 The Mathematical Theory of Communication. *Bell System Technical Journal* 27: 398-403.

Sinclair, John M.

1987 [Ed.] *Looking Up—An Account of the COBUILD Project in Lexical Computing*. London: Collins.

1991 *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sproat, Richard, and Chilin Shih.

1990 A Statistical Method for Finding Word Boundaries in Chinese Text. *Computer Processing of Chinese and Oriental Languages* 4.4: 336-351.

Svartvik, Jan.

1992 [Ed.] *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, 4-8 August 1991. Trends in Linguistics Studies and Monographs 65*. Berlin: Mouton.

Wang, William S.-Y.

1988 Diannao zai Yuyanxue li de Yunyong [The Use of Computer in Linguistics]. Keh-jiann Chen and Chu-Ren Huang Eds. *Proceedings of R.O.C. Computational Linguistics Workshops I (ROCLING I)*. 257-287. Taipei: Computational Linguistics Society of R.O.C.

Zampoli, Antonio, Christain Boitet, Nicoletta Calzolari, and Sergio Rossi.

1992 [Eds.] *The Proceedings of the 1992 International Conference on Computational Linguistics (COLING-92)*. Nantes, France.

Zhou, Xiaolin, Ruth Ostrin and Lorraine Tyler.

1993 The Noun-Verb Problem and Chinese Aphasia: Comments on Bates et al. (1991). *Brain and Language* 45: 86-93.

Chu-Ren Huang, Kathleen Ahrens & Keh-jiann Chen

由語料出發驗證心理詞庫——漢語語料庫語言學研究二例

黃居仁／安可思／陳克健

中央研究院語言所／國立台灣大學外文系／中央研究院資訊所

本文試圖由語料著手來探索語言之心理真實性。傳統研究是以實驗爲依據。這類心理或腦神經語言學研究雖然得到了不少突破。但仍有其限制。首先實驗室迫使受試者在受控制的非自然環境中使用語言；其次實驗的設計往往只限於少數幾個句子；最後限於受試者注意力的限制，實驗語句限制長度而缺乏自然的上下文語境。本文認爲大量語料除可補足上述實驗方法之不足，且可表現出語言的心理真實性。

以語料庫探索心理真實性的前提有三：一、語料庫提供了在自然環境下語言使用（生成）的實例。二、語料庫正好也代表了日常語言辨識對象的大量取樣。三、適當抽取的語料正可以呈現使用該語言的人所共有的語法知識。

文中討論了兩個研究，這兩個研究均是根據中央研究院現代漢語語料庫爲基礎。第一個研究探討中文的複合詞，第二個研究探討中文特殊的構詞現象——「縮寫」。這兩個研究都支持了一個基本假設——即「詞」這個觀念在漢語的心理詞彙庫中的確存在而且可以利用語料庫資料判讀。也就是說語料庫反映了語言的心理現象，可提供了我們由資料入手研究語言真實性的另一蹊徑。

關鍵詞：心理詞庫 詞 語料庫 互見訊息