In Mineharu Nakayama (Ed.) Sentence Processing in East Asian Languages.
CSLI Lecture Notes. Stanford: CSLI Publications.

# The Nature of Categorical Ambiguity and Its Implications for Language Processing: A Corpus-based Study of Mandarin Chinese[1]

CHU-REN HUANG, CHAO-JAN CHEN AND CLAUDE C.C. SHEN
*Academia Sinica*

Ambiguity is a 'design characteristic' of the human cognitive system, even though scientific endeavors generally regard ambiguity as an evil that must be avoided at all cost in a representation system. Ambiguities are also costly to resolve in any psychological or computational model of language processing (e.g. Small et al.1988, Gorfein 1989, Schutz 1997, Ahrens 1998). However, the understanding of ambiguity may hold the key to the understanding of human cognition.

In this paper, we take an empiricist approach towards a descriptive account of the nature of categorical ambiguity in Chinese. The two assumptions of this study are that i) categorical ambiguity is lexical by nature, i.e. a word is categorically ambiguous only when it is assigned multiple categories in the lexicon; and that ii) lexical representation of grammatical categories must be attested by actual use.

Based on the above assumptions and given the fact that it is still impossible to enumerate contents of any individual mental lexicon (Huang 1995), we base our exploration on the information of an annotated corpus: the 5 million word segmented and tagged Sinica Corpus (Chen et al. 1996). A lexicon of over 14,000 entries is compiled based on the corpus. Any entries that are assigned

[1] The authors would like to thank Kathleen Ahrens for her helpful comments on various versions of this paper. The comments of the participants of the 1999 International East Asian Psycholinguistics Workshop as well as colleagues at CKIP, Academia Sinica, are appreciated. The research conducted here was partially supported by grant 89-2420-H-001-002 from the National Science Council of Taiwan, ROC. Any remaining errors are, of course, our own responsibility.

more than one category in the corpus are considered categorically ambiguous. We will answer the following questions based on the data:

1) What is the degree of categorical ambiguity in a natural running text?
2) Will the size of category sets affect the degree of ambiguity?
3) Are some natural classes more ambiguous than others (Gentner 1981)?
4) Is there always a strong default in categorical ambiguities?

Results reported here will have important implications for both computational and psycho-linguistics, as well as the fundamental issue of knowledge representation in cognitive science.

## 1. Background

### 1.1 Premises

We will show in this paper that corpus-based approaches are highly relevant to the study of human language processing and have important insights to offer to its theory. In particular, we will show that corpus-based approaches allow us to study ambiguity globally and systematically as an integral part of human grammatical representation. Such approaches, we claim, lead to better understanding of the nature of ambiguity as well as of the human linguistic system.

Our three premises are that:

1. ambiguity is a 'design characteristic' of the human cognitive system;
2. corpora offer reliable empirical data of active ambiguity in human language processing; and
3. ambiguity must be introduced lexically.

First, resolution of ambiguity is a recurring topic in studies of both psychological and computational language processing (e.g. Small et al. 1988, Schutz 1997). It is also a central issue in linguistic semantics (e.g. Lyons 1977, and Pustejovsky 1995). In contrast, artificial languages, including those intended to be spoken by people, tend to be ambiguity-free. To truly understand human cognitive system, we feel that it is inadequate to take discovery and accounting of strategies or algorithms of ambiguity resolution as theoretical aims. If the observation that ambiguity is one of the defining features of human grammatical representation is correct, then it is theoretically more pertinent for language processing studies to account for why ambiguity exists rather than how to resolve it. And of course, knowing the 'why's' should help us to implement the 'how's' in a principled way with better results.

Second, if ambiguity is taken to be a design characteristic of human language, then it is the role that ambiguity plays in a grammatical system that is critical to the theory of language and cognition. In other words, we want to know what conceptual or representational issues in human language and cognition induce ambiguity in grammar. To answer this essential question, we must study a complete grammar and make generalizations over all possible ambiguities. However, current psycholinguistic experiments allow us to study ambiguity only locally. In addition, since it is not possible to describe directly

and completely the mental grammar of any given person, it is also not possible to extract formal properties from such grammatical systems.

Huang (1994) argues that 'a grammar of language X' cannot be defined in the Chomskyan sense as the final mental states of individual language acquisition. Adopting Chomsky's (1995) UG-type definition, assuming that speaker of the same language share a grammar, it leads to the conclusion that a child acquires the grammar of his/her parents. However, since our parents acquired the grammar of their parents, and so on so forth, the argument implies that contemporary Chinese share the same grammar as first century (or earlier) Archaic Chinese, which is a simple and straightforward fallacy.

On the other hand, if each individual mental grammar is taken to be different, there is still need to describe the 'sameness' of the set of 'grammars' that form the 'language' as shared by a given speaker community, such as English or Chinese. If this 'sameness' is to be defined according the Chomskyan mental grammar, then we will lead to the same fallacy as above.[2]

Huang (1994) concludes that '(t)he grammar of any given language L is the set of shared (linguistic) knowledge of the speakers of L.' If this position is adopted to account for the notion of 'the grammar of language L,' then the only realistic alternative to study grammar of a language L is to study a comprehensive subset of actual language use shared by speakers of a language. We claim that corpora offer such a fragment. In particular, since a balanced corpus contains language production produced by a wide range of speakers and accepted for comprehension by the majority of speakers, we assume that it contains reliable empirical data of active ambiguity in the language. By 'active' ambiguity, we refer to the ambiguities that are instantiated in language uses. In other words, study of ambiguities in a corpus allows us to approximate the nature of ambiguity in a grammatical system.

Third, in order to study ambiguity, we need to know the sources of ambiguity. Although linguists talk about different types of ambiguity such as lexical ambiguity, structural ambiguity, and contextual ambiguity, it is a simple fact of the grammar that no meaning can be accessed unless it is represented in the mental lexicon. Even for the so-called structural ambiguity (e.g. PP attachment), it is a pre-requisite that the words involved in the structures are lexically encoded to allow either interpretation.[3] In other words, it is the lexicon that provides the full-range of ambiguities. The ambiguity-inducing contexts (lexical collocation, structure, discourse contexts etc.) actually restrict ambiguity to certain alternatives. To sum up, in accordance with a lexicalist and modular

---

[2] Take note that this cannot be accounted for by invoking the langue/language distinction of Saussure. If the distinction were between individual language use and the system of langue, than we will induce that there is only one langue for all human beings, again contradicting Saussure's original definition of langue as the grammatical system shared by speakers of a given language (i.e. English or French).

[3] It is worth noting that, similar to Pustejovsky (1995) we take ambiguity to be a representational and interpretational issue, not a processual issue. In other words, a linguistic item is only ambiguous when its representation allows multiple interpretations. Constructions such as garden path sentences allows only one interpretation. Thus, while being misleading and costly processing-wise, they are not ambiguous.

view of grammar, ambiguity is representational in nature and must be introduced lexically.

### 1.2 Goals

In this study, we concentrate on categorical ambiguity. This is because 1) grammatical categories are better defined and 2) category-tagged corpora are available for study. Based on the above premises, we identify the goals for this study as: First, to establish the distribution of categorical ambiguity in the language. This includes how often categorical ambiguity occurs in the language, as well as the distribution of the categories involved when ambiguity does occur. One of the expected results of this study will be the baseline performance benchmark of categorical ambiguity resolution. In practice, we will take the expected performance of ambiguity resolution with lexical information alone as the baseline. This benchmark can be used then to measure the success of a computational category tagging or ambiguity resolution algorithm. It can also be used as a parameter to measure whether an ambiguity resolution task is successfully completed in a human language processing experiment. Second, in order to better understand the nature of categorical ambiguity in human languages, we will extract generalizations involving categorical ambiguity. In particular we would like to establish correlations between categorical ambiguity data and lexical factors. For instance, we will look into the correlation between categorically ambiguity and frequency, categories, syllabic length, etc. Third, our ultimate goal is to model the lexical representation of categorical ambiguity. Our first steps towards a theoretical model will include the description and account of default category, as well as a comparative study of the effect of the size of the categorical set on degree of ambiguity.

## 2. Methodology

As mentioned above, since it is not possible to examine a complete individual mental lexicon, we extract our data from a corpus instead. A corpus is treated as collective language use of the population for the following two reasons. First, a corpus is the collection of the language production of all the authors/speakers involved. And second, a corpus, especially a balanced one, is a representative sample of language comprehension data received by the majority of the speakers. In the first sense, a corpus is the instantiation of the collection of individual active lexicons. In the second interpretation, a corpus is the instantiation of a shared passive lexicon of the language. In this study of categorical ambiguity, we assume the second interpretation in our account.

As a logical consequence to our corpus-based approach, we take an empirical definition of categorical ambiguity. In other words, all and only forms which have attested instantiation with more than one category in the corpus are considered categorically ambiguous. While we cannot rule out the possibility of un-instantiated lexical ambiguity, such rare un-instantiated instances will have at most a marginal effect on the result of our statistical characterization of the entire corpus.

Based on the above assumption, each instance of the categorically

ambiguous lexical item in the corpus represents an instance of successful resolution of the lexical ambiguity. Thus, our task is straightforward: To examine all such instances in the corpus, and to extract generalizations that will shed light on the nature of categorical ambiguity as well as its interaction with other grammatical elements.

**2.1 Data**

The corpus we use is Sinica Corpus 3.0., i.e. version 3.0. of the Academia Sinica Balanced Corpus for Modern Mandarin (Chen et al. 1996). It was completed in 1995 and contains 5 million words of modern Taiwan Mandarin. The data includes mostly written texts as well as a small portion of transcribed spoken texts. These texts are mostly from the 80's and early 90's. Each text is segmented to mark word boundary and each word is tagged with grammatical category. The segmentation and tagging were machine-aided. In other words, automatic segmentation and tagging were performed by machine to give human taggers the raw material to work with. The human taggers then either accept the machine-given default or assign a new boundary/category according to his/her analysis. Each tagged and segmented text is crossed-checked by both human and machine to eliminate inconsistency and human error. In addition to local versions, Sinica Corpus is available on the web for search and use:

http://www.sinica.edu.tw/ftms-bin/kiwi.sh

**2.2 Tagset: The Set of Grammatical Categories Chosen**

There are forty-six (46) categories used in the Sinica Corpus for tagging. This category set is a modified version based on Chao (1968), and described in CKIP (1995). The number of categories is of the same scale as many modern day corpora in other languages. Hence it allows all together 2,070 (=46x45) possible pairs of two-way ambiguity. However, only 1,375 pairs are attested in the corpus. The number of non-attested pairs is 695. We will be looking into these pairs in the future to decide which of them are genuinely impossible, as well as if there is any linguistic motivation or explanation for the absence of such types of ambiguity.

In addition, when the Sinica Corpus data was used in constructing a digital museum/library site, the category set was reduced to thirteen (13) for the general public that we expect to reach. Linguists as well as elementary school teachers were involved in deciding the set of 13 categories such that they are both linguistically felicitous and pedagogically intuitive (Huang 1999). The 13 generalized categories and their corresponding sub-categories are listed below in Table 1. Please see Chen et al. (1996), as well as online help at the Sinica Corpus website, for a definition of the complete category set[4].

[4] FW (Foreign Word) is theoretically speaking not a grammatical category. However, it is a necessary corpus tag especially in this age of high tolerance of multi-lingualism and frequent code-switching. The generalized category set will contain 12 categories, and the complete set 45 categories, if FW is not counted.

**(1) Generalized and Complete Category Sets for Sinica Corpus**

| | Generalized Tag | Category Name | Equivalent Complete Categories |
|---|---|---|---|
| 1. | A | Adjective | A |
| 2. | C | Conjunction | Caa, Cbb |
| 3. | ADV | Adverb | Da, Dfa, Dfb, Dk, D |
| 4. | ASP | Aspect (marker) | Di |
| 5. | N | Noun | Na, Nb, Nc, Ncd, Nd, Nh |
| 6. | DET | Determiner | Neu, Nes, Nep, Neqa |
| 7. | M | Measure (word) | Nf |
| 8. | T | Particle | I, T, DE |
| 9. | P | Preposition | P |
| 10. | Vi | Intransitive Verb | VH, VA, VB, VI |
| 11. | Vt | Transitive Verb | VAC, VC, VCL, VD, VE, VF, VG, SHI, VHC, VJ, VK, VL, V-2 |
| 12. | POST | Postposition | Ng, Neqb, Cab, Cba |
| 13. | FW | Foreign Word | FW |

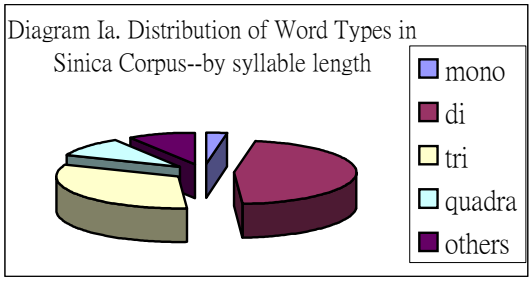
The variation in the categorical sets brings up the issue of the psychological reality of sets of grammatical categories. While the psychological reality of a single category is easy to justify with either distributional/linguistic data or with psychological experiments, it is far more difficult to justify a complete system of categorization, such as the system of grammatical categories. We are not aware of any previous empirical attempt to justify the choice of a system of grammatical categories. The usual practice is simply to adopt an established linguistic account. However, we observe that corpus and computational linguistics often adopts a category set much larger than that adopted by psycholinguistic work. The size that a computational work adopts varies from 40 to 100. For instance, the British National Corpus has sixty-five parts of speech, while a recent French corpus under construction adopts over forty categories (Abeille 1998). On the other hand, a typical category set a psycholinguistic work adopts is between ten and twenty. For instance, Redington et al. (1995) justified their reduction of the Sinica Corpus category set to 11 by mentioning that the typical number of grammatical categories used in psycholinguistics is less than a score.

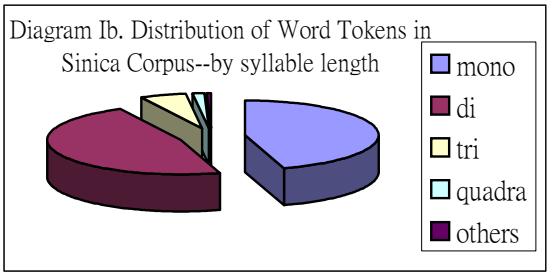
While it is beyond the scope of this current study to test for the validity of a given system of grammatical categories, we would still like to shed some light on the observed disparity of categorical set size chosen by psychological and computational linguists. Take Redington et al. (1995) for example again, the paper and other related simulation work proved that a simple distributional learning mechanism can acquire grammatical categories from un-analyzed linguistic data. However, since no categorical set has been independently verified as psychologically real and as the goal of language acquisition, it is justifiable to ask if the same result can be obtained when a different categorical set is adopt, especially one that differ substantially in scale. In this current study of categorical ambiguity, we will compare the data involving both the

generalized and the complete category sets of Sinica Corpus. Since these two sets are both linguistically well-motivated and one subsumes the other, it can be argued that their only difference is in their size. Thus, the result of the comparative study will tell us if the choice of size of category set affects the nature of categorical ambiguity.

**2.3 Basic Distribution**

To sum up the introductory section, we would like to clarify the basic facts regarding syllabification of Mandarin Chinese. On one hand, Chinese is often claimed to be a monosyllabic isolating language. Even though this may be historically true, the modern language is definitely not monosyllabic. On the other hand, it has also been repeatedly claimed recently that Mandarin Chinese is undergoing di-syllabification and disyllabic words are dominant in modern usage. This also proves to be over-exaggeration. Our complete statistics of token and type distribution of Sinica Corpus shows that while disyllabic words dominate the lexicon (i.e. type distribution), they do not exceed 50% (Diagram I). While disyllabic words take up 46.06% of all lexical entries, monosyllabic words take up a mere 2.73%.



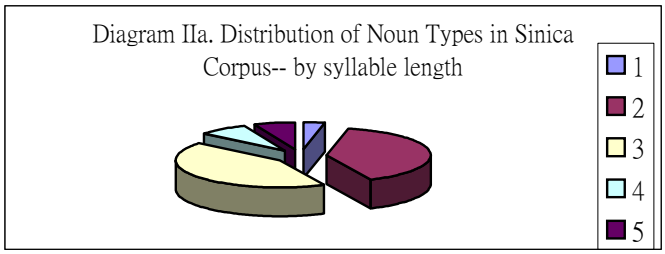

| **Syllabicity** | **Type** | **Token** |
|---|---|---|
| monosyllabic | 2.73% | 45.33% |
| disyllabic | 46.05% | 47.25% |
| tri-syllabic | 32.41% | 5.57% |

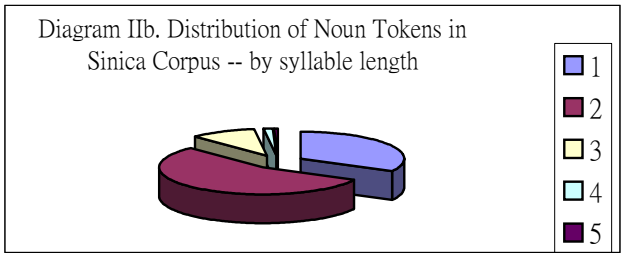
| | | |
|---|---|---|
| quadra-syllabic | 10.05% | 1.29% |
| all others | 8.75% | 0.56% |

Diagram I also shows that while mono-syllabic words takes up a small percentage in the lexicon, they tend to be highly frequent and occur almost as frequently as disyllabic words as a group in actual use. Total occurrences (i.e. token frequency) of monosyllabic and disyllabic words are roughly equivalent, with disyllabic words at 46.83% and monosyllabic words at 45.83%. The above corpus data can be used to explain the conflicting claims of mono- and di-syllabicity of Chinese. The highly frequent use of monosyllabic works enhances the cognitive saliency of monosyllabic words in Chinese[5]. However, on the other hand, the large number of disyllabic entries leads to the impression that Mandarin is a disyllabic language. While the truth is that these two tendencies are both at work and dominant at the type and token levels respectively. Words of higher syllabicity, however, do occur and represent substantial fragments of the language.

註解 [Chu-Ren1]:

Distribution of nouns and verbs are given in diagram II and diagram III respectively to show that syllabic distribution varies from category to category.

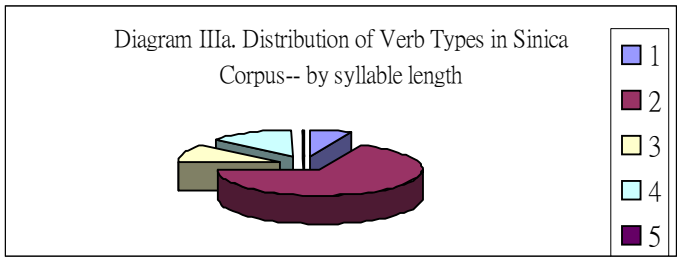


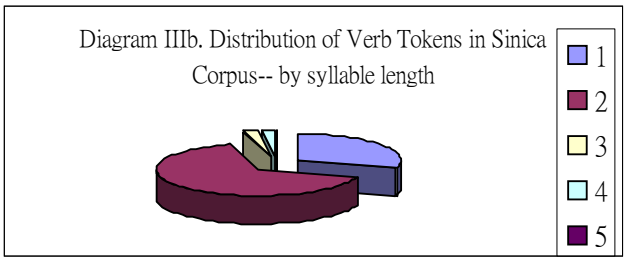




**Syllabicity** **Type** **Token**

[5] Eighty-six of the one hundred most frequent words in this corpus are mono-syllabic, the other fourteen are di-syllabic. And the most frequent di-syllabic word was ranked eighteenth overall (Huang et al. 1998c).

| | | |
|---|---|---|
| monosyllabic | 3.30% | 32.84% |
| disyllabic | 39.67% | 55.03% |
| tri-syllabic | 42.97% | 10.17% |
| quadra-syllabic | 7.79% | 1.37% |
| all others | 6.27% | 0.59% |




| **Syllabicity** | **Type** | **Token** |
|---|---|---|
| monosyllabic | 6.72% | 29.36% |
| disyllabic | 67.33% | 66.10% |
| tri-syllabic | 12.22% | 2.39% |
| quadra-syllabic | 13.42% | 2.13% |
| all others | 0.31% | 0.02% |

The above diagrams show that verbs and nouns have different distributional properties based on their syllable length. Although verb types are predominantly disyllabic (over 67%), there are actually more noun types which are tri-syllabic (almost 43%) than disyllabic (close to 40%). This can be explained by the pragmatic fact that Chinese names are predominantly tri-syllabic (monosyllabic family plus disyllabic given name). It is also worthwhile to point out that even though monosyllabic words represent only a small fraction of the verbal and nominal entries, they are used more frequently than any other syllabic type. For instance, the token distribution of monosyllabic

nouns is 10 times bigger than their type distribution. This means that monosyllabic nouns have an average frequency ten times bigger than the average frequency of all nouns. Similarly, monosyllabic verbs are more than four times as likely to be used than an average verb. In contrast, disyllabic nouns and verbs have a close to average distribution, while all the other longer syllabic types occur much less often than average. One last observation involves the quadra-syllabic verbs. Although lexical types and tokens in Chinese tend to become less frequent as syllabic length increase, quadra-syllabic verbs are actually slightly more than tri-syllabic ones in terms of types and roughly equivalent in terms of tokens. This can be attributed to the popularity of Cheng2yu3 , four syllable idiom chunks, in Chinese. Cheng2yu3 are typically used as an intransitive predicate in Chinese.

## 3. Categorical Ambiguity and Distributional Factors

### 3.1 Degree of Categorical Ambiguity

Only 4.298% (6316/146,929) of all the lexical entries represented in the corpus are assigned with more than one category. In other words, these are the words that call for categorical ambiguity resolution when they are encountered in processing. However, the total occurrences of these lexical entries take up 54.59% of the corpus. In other words, even though only a small portion of the lexicon is categorically ambiguous, their usage represents more than half of the corpus. In natural language processing terms, the small number of lexical items involved means that the lexical knowledge of categorical ambiguity can be dealt with exhaustively and heuristically. In addition, regarding on-line ambiguity resolution, the high percentage of potential ambiguity implies that human resolution of categorical ambiguity is highly principled and efficient, since human language processing would be prone to mistakes and highly inaccurate otherwise. This fact also suggests that there is a high correlation between frequency and ambiguity, which will be looked at in more details in the next section.

Another way to look at degree of categorical ambiguity is to find out the distribution of the number of possible categories for all ambiguous words. We found that of all ambiguous lexical types, two-way ambiguous types dominate and represent roughly half of all types (49.478%). In addition, three-way ambiguous types represent 9.832% of all types, and percentage of higher multi-way ambiguity gets progressively lower. Because of the near majority of the two-way ambiguous types, and because all higher order ambiguity can be factorized into several two-way ambiguous types, our current study will focus on binary contrasts as the basic type.

### 3.2 Categorical Ambiguity and Lexical Frequency

We observed in the last section that categorically ambiguous words represent only a small portion of the lexicon but take up more than half of the corpus in actual use. To illustrate a possible correlation between frequency and categorical ambiguity, complete data from the 5 million words corpus is used to plot Diagram IV. Degree of ambiguity is calculated according to intervals of

frequency ranking. For instance, 320 of the 500 most frequent words are categorically ambiguous, thus the degree of ambiguity of this interval is 64%; while only 62 of the 500 words ranked between 10,001 and 10,500 are ambiguous, with a degree of ambiguity of 12.4%.[6]

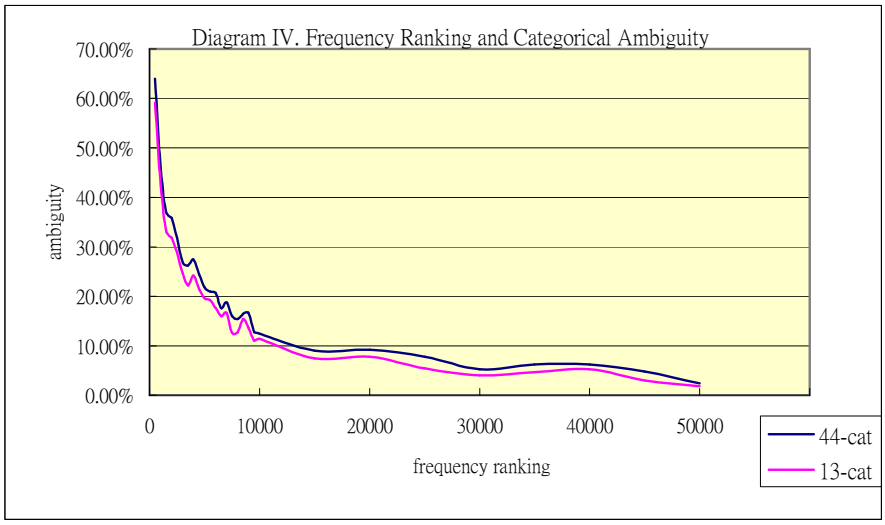



Diagram IV shows that there is a high correlation between frequency ranking and categorical ambiguity. This diagram also shows distinct areas of correspondence: First, from the top rank to the rank around 1,000, the correlation is highly regular and ambiguity level descends acutely. Second, between the rank of roughly 1,000 to 10,000, the decrease is less acute and seems to fluctuate more. Lastly, for the rank above 10,000, the degree of ambiguity is so low that the frequency ranking does not seem to have any significant effect.

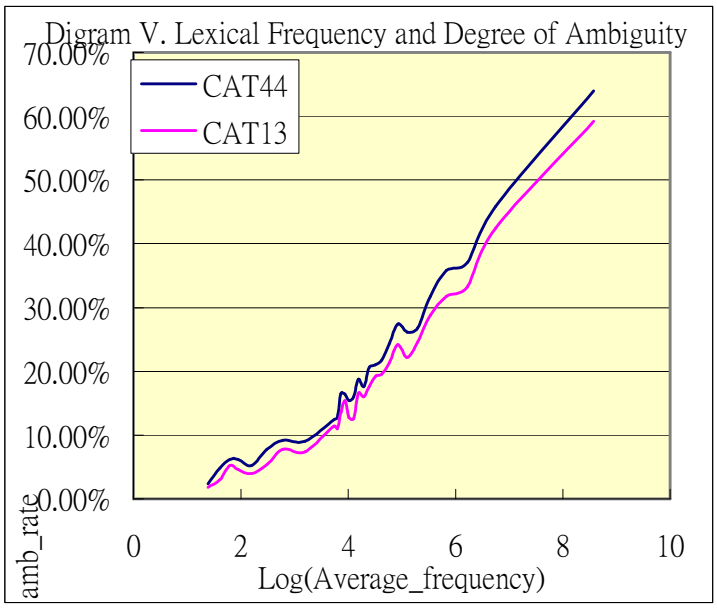
It is interesting to note that these three areas seem to correspond to commonly used criteria to define 'frequent', 'less frequent,' and 'infrequent' lexical entries. In Huang et al.'s (1998c) statistics based on Sinica Corpus, the accumulated frequency of the top 456 words just surpasses 50%, the top 1,000 words cover nearly 60%, while the top 10,000 words covers 85.87%. In terms of lexical frequency, the 456$^{th}$ word has a frequency of 0.0259%, the 1,000$^{th}$ word 0.0128%, and the 10,000$^{th}$ word 0.000896%.

Theoretically, however, we need to measure degree of ambiguity directly against lexical frequency in order to understand the nature of the relation

---

[6] In this section, we will restrict our discussion to the data involving either the 13 or the complete 44 categories, although the data are sometimes shown together. The contrast between the complete and generalized 13 categories will be discussed later in Section 5.

between frequency and ambiguity. The above result is inadequate since frequency ranking is only a second order representation of lexical frequency. Methodologically, however, we have a problem to overcome before the frequency-ambiguity relation can be studies. Note that lexical frequency is the attribute of an individual lexical item and that each lexical item is either ambiguous or unambiguous. In other words, degree of ambiguity of a lexical item is either 1 or 0. In order to have a degree of ambiguity data, we need to define a sub-set of lexical items that share similar frequency features and calculate the number of ambiguous words among them. A significant way to achieve this is to group lexical items according to their frequency ranking. Thus lexical items in each group will share similar lexical frequency, while their frequency will differ significantly from other lexical items outside of the group. Hence it will be justified to take the average frequency of the whole group as their shared lexical frequency, and to examine its correspondence to the degree of ambiguity of the group. One further technical detail is that logarithmic representation of lexical frequency will be adopted, following well-established tradition to better demonstrate frequency variations. Diagram V is the result of plotting degree of ambiguity against their average frequency for each 500 lexical items arranged according to descending frequency rank (i.e. from the least frequent to the most frequent).



The correlation between lexical frequency and degree of ambiguity is succinctly represented by Diagram V. The diagram shows unequivocally that there is a direct proportional relationship between frequency and degree of ambiguity. The more frequent a lexical item is, the more likely for it to be categorically ambiguous. This is another instantiation of Zipf's law (see Manning and Schutz 1999 for an interpretation.)

The evolutionary view on language (Cavalli-Sforza 1994, Cavalli-Sforza and Wang 1986) offers a possible explanation of the above correlation between frequency and categorical ambiguity. It is widely accepted that gene mutations are highly correlated to the rate of its reproduction. In other words, the more often a gene is copied, the more likely it is to mutate.

With regard to linguistic categories, we can treat each use of a lexical item in a context similarly to an instance of gene copying. In other words, each is an instance of instantiation of an archetype. Hence, as the frequency of use of a lexical item rises, so does its chance to change, including to a shift in its grammatical category. Of course, frequency of use does not guarantee change, just as frequent reproduction does not guarantee mutation either. It is simply that the statistical probability rises significantly. We expect this account to receive future support from the pioneering research on the evolution of language.[7]

### 3.3 Grammatical Categories and Categorical Ambiguity

An interesting theoretical question to ask is if ambiguity is category-dependent. If words belonging to a certain grammatical category are more ambiguous than others, then it is natural to follow up with the question of whether the defining features of that particular category directly lead to its tendency for categorical shift. Hence study on categorical dependency of ambiguity could offer crucial evidence for human cognitive system; particularly if it is shown that the propensity for ambiguity is driven by the same cognitive/conceptual characteristics defining a category. Initial data that degree of ambiguity do range widely (from just over 5% to over 60%) by categories seem to suggest a correlation between categorical ambiguity and categorical classification, as listed below in Table 2:

**(2) Degree of Ambiguity for Lexical Categories**

| Category | Amb. Entry/Total Entry | Percentage |
|---|---|---|
| Noun | 4727/93295 | 5.067% |
| Verb | 3939/44241 | 8.904% |
| Adjective | 430/1912 | 22.490% |
| Interjection | 22/76 | 28.974% |
| Adverb | 935/2664 | 35.098% |
| Preposition | 180/306 | 58.824% |
| Particle | 48/78 | 61.538% |

[7] Note that this account is also compatible with the lexical diffusion theory for language change and language variation (Wang 1969, 1991).

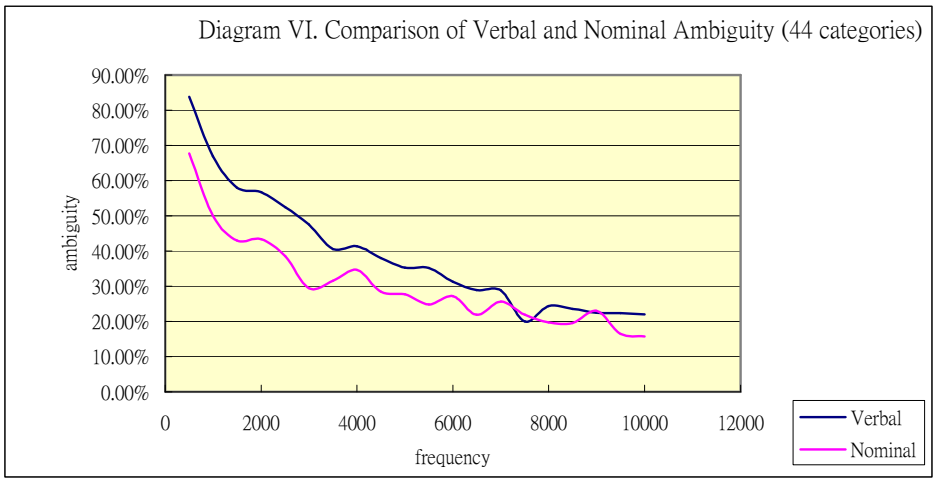
However, there are two theoretical issues which indicate that the above data alone are not sufficient to lead to the conclusion that ambiguity is category-dependent. The first involves the assignment of a 'primary' category to a lexical item. The identification of a primary category is crucial in order to establish the correlation between category and ambiguity, since all the categorically ambiguous lexical items are assigned with more than one category in the corpus and lexicon. Is there a general and principled way to assign the primary categorical affiliation of a categorically ambiguous word and hence attributes ambiguity to that category? The answer is a qualified no. Etymology and speaker perception are two obvious ways to determine primary category for a lexical entry. However, based on past experience as well as our pilot analyses, they often conflict with each other and there do not seem to be any principled ways to resolve the conflicts. On the other hand, even if a 'primary' category can be assigned with consensus, it is still not valid to make the deduction that the lexical categorical ambiguity be attribute to it. One could as easily argue that a lexical item shifts to a secondary category simply because it has some of the characteristics of that category. Thus, can the categorical ambiguity could be attributed to the characteristics of the secondary (or tertiary, etc.) category.

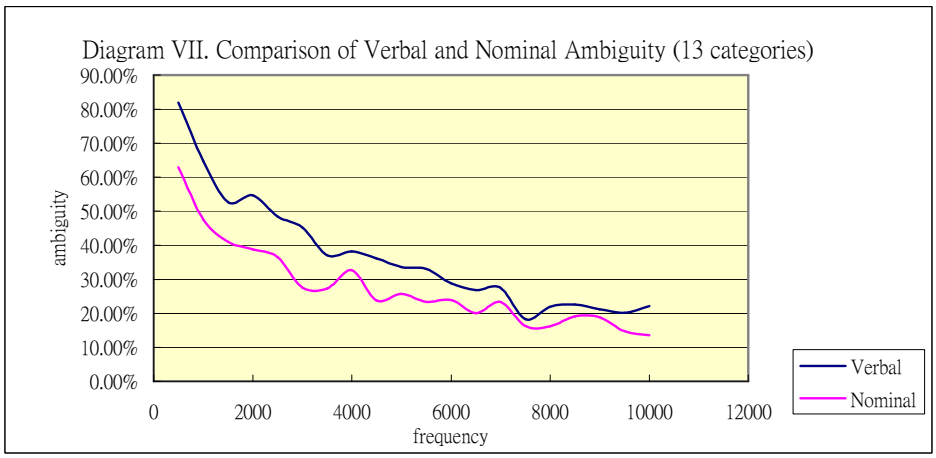
To overcome the above potential problems, we take a naïve but non-discriminatory approach for the current study. That is, any lexical item that is assigned multiple categories is counted towards the statistics of each and every category it is assigned to. In other words, if a word is assigned 3 possible categories, than it is added to the ambiguous word list for all the three categories. This way, we can avoid making arbitrary decisions to rule out the contribution of any category. And since all attested categories are counted, there is also no danger of any specific category being given privilege over others.

The second problem is that, as shown in the last section, ambiguity is highly dependent on frequency. It is also known that lexical frequency varies greatly from category to category. For instance, the closed categories tend to have a smaller number of members that are more frequently used. Hence the claim that such categories are more ambiguous may be misleading since their ambiguity can actually be attributed to the frequency effect. To overcome the possible influence of frequency effect and yet investigate the correlation between category and categorical ambiguity, we need to compare two categories that are versatile and have the same wide range of distribution. Two of the categories that meet these requirements are verbs and nouns.

It is interesting to note that nouns and verbs happen to be at the center of dispute over the so-called mutability issues. Genter (1981) first observed that verbs are consistently more polysemous than verbs across different frequency ranges in English. Genter and France (1988) offered the account that verbs are more mutable than nouns. This claim was challenged by S. Huang (1995) and received a qualified support and revisions from Ahrens (1999), both based on Mandarin Chinese data. Genter and France's (1985) original cognitive claim was that verbs are less concrete and therefore more susceptible to change. S. Huang's (1995) work based on counting sense entries in a dictionary raises the issue that this motivation may be language-dependent. Ahrens's (1999) off-line

experiments on native speakers are more strictly controlled regarding syllable length and frequency. However, none of the above studies were able to an across-the-board look of the data from a language. The Sinica Corpus data offers a chance for us to have a comprehensive look of contrasts between verbs and nouns across all frequency ranges.

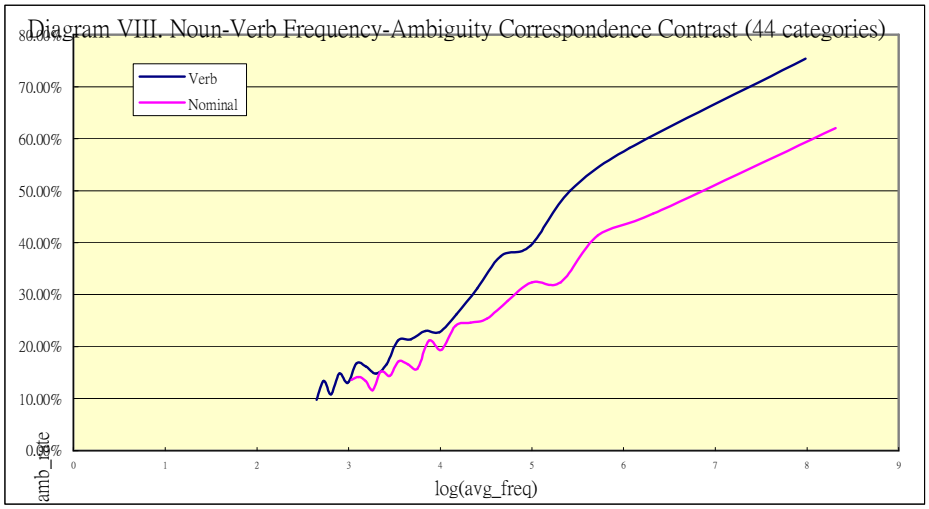


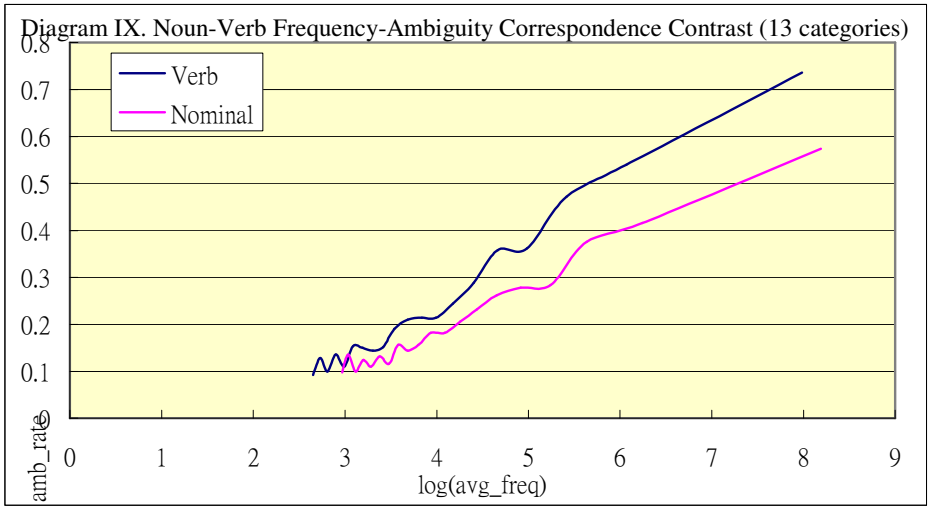

In Diagrams VI & VII above, we see a consistent gap between the degree of ambiguity between nouns and verbs at every frequency range. Thus verbs are more likely to be categorically ambiguous than nouns. An interesting

observation can also be made when comparing diagrams VI and VII with diagram IV. It is shown that the degree of ambiguity of both nouns and verbs are higher than the average degree of ambiguity of all categories. In other words, these two 'substantive' categories are more likely to be categorically ambiguous than others. This actually supports Ahrens's (1999) position that verbs and nouns are both mutable but for different reasons.[8] Thus, we offered empirical evidence to show that categorical ambiguity is indeed dependent on categorical identity. Whether such dependency can be shown to be driven by the same conceptual or cognitive motivations for categorical classification will be an interesting research topic for future study.

Recalling our discussion on frequency ranking and lexical frequency earlier, even though the above frequency ranking data give us good indication that verbs are more likely to be categorically ambiguous than nouns, it is more interesting to show direct correspondence between lexical frequency and degree of ambiguity. In the current study, the danger of distortion is even greater since verbs and nouns may differ in individual frequency, even when they belong to the same frequency range. Following the methodology established earlier, we plot (logarithmic) average frequency against their degree of ambiguity according to frequency ranking range of every one hundreds words, for both nouns and verbs.




[8] The proposal is that a verb is more likely to change in situations when the change is not crucial to its meaning. On the other hand, a noun is more likely to change in a situation when the change brings a more specific interpretation.




Diagrams VIII and IX both show two important generalizations. First, they show that the direct proportional relationship between frequency and degree of ambiguity holds for individual categories too (i.e. for nouns and verbs). Second, They show unequivocally that verbs are more likely to be categorically ambiguous than nouns given that they have the same frequency. In addition, the diagrams also show that the categorical ambiguity disparity between verbs and nouns grows when frequency increases.

## 4. Lexical Knowledge and the Resolution of Categorical Ambiguity

In this section, we will explore the role lexical knowledge plays in the resolution of categorical ambiguity. In particular, we would like to find out the distribution of all possible categories for a lexical form. If the distribution is somewhat random, then categorical ambiguity is not constrained by lexical knowledge and is totally dependent on contextual rules. On the other hand, however, if generalizations on categorical distribution can be extracted for all lexical entries, than how to represent such generalizations in the lexicon becomes a crucial issue.

In a model where all possible categories of a ambiguous word are accessed with equal probability, one would predict that, barring contextual coercion, all possible categories will have similar frequency. It would also predict that the difficulty of ambiguity resolution of any lexical entry is directly proportional to the number of available categorical choices. Neither of the above predictions bore out. A look at the data shows that 1) Most lexical forms have a dominant category in terms of frequency of use, as will be discussed later in this section. 2) Assuming the modular view that language production and comprehension

presupposed correct assignment of grammatical categories to all lexical items, native speakers rarely, if ever, makes mistakes in category assignment.

An alternative model, without having to resort to the yet controversial issue of lexical encoding of stochastic information, is where a default category can be specified for each lexical entry. Lexical default ensures that a category is assigned in a principled way even when there is insufficient contextual information. Such a default mechanism will allow the speaker to beat the statistical odds without going through complex calculation. Such a model will also provide the baseline performance standard for computational tagging (i.e. categorical assignments), to be discussed in the next sub-section.

**4.1 Categorical Default**

In this study, we follow the standard practice and lexically encode the most frequently instantiated category of each word form as its default category. We assume that a speaker's null hypothesis is to assign the default category to each instance of the word form unless context information indicates otherwise. In this model, categorical assignment and ambiguity resolution does not start with either zero or statistical ambivalence. It starts with all the correct prediction based on lexical knowledge. Lexical knowledge can provide correct categorical assignments in two ways: First, it automatically assigns the correct category when a lexical item is unambiguous. This happens to nearly half of the tokens in Sinica Corpus. Second, when there is categorical ambiguity, lexical knowledge makes the correct prediction when the actual category is the lexical default. The Sinica Corpus data also shows that lexical default holds true for 88.37% over all ambiguous tokens. Taking the two scenarios together, a purely lexical approach to categorical ambiguity resolution and assignment will have very good results according to our corpus data. The frequencies of the default category being instantiated are shown in Table 3 according to both syllable length and number of possible categories.

**(3) Frequency of Default Category**

| Syllable Number | Freq. (by type) | Freq. (by token) |
|---|---|---|
| 1 | 80.21% | 88.86% |
| 2 | 78.34% | 86.87% |
| 3 | 73.73% | 89.70% |
| 4 | 69.98% | 81.82% |
| 5 | 69.02% | 80.02% |
| 6 | 65.01% | 74.35% |
| 7 | 65.15% | 83.33% |

| No. of Categories | Freq. (by type) | Freq. (by token) |
|---|---|---|
| 2 | 77.65% | 91.21% |
| 3 | 77.71% | 88.39% |
| 4 | 74.21% | 89.50% |
| 5 | 73.83% | 92.43% |
| 6 | 73.46% | 86.09% |
| 7 | 68.51% | 86.09% |
| Total | | 77.36% |

MEAN = 77.3556%
DEV = 17.2175%

However, it is interesting to observe that if the precision rate of assigning the default category is calculated for each lexical form, and the average precision rate of all lexical types is actually only 77.36%, much lower than the overall precision rate. The only possible explanation for such discrepancy is that the default precision rate for more frequent lexical types is higher than less frequent ones.

Another observation of Table 3 is that default categorical precision does not seem to correlate with either syllable length or number of possible categories. We will discuss the implications of the two observations at the conclusion section.

**4.2 Estimating the Baseline Performance Categorical Ambiguity Resolution**

Given the model of lexical assignment of default categories as well as the data from Sinica Corpus, we will construct a baseline model of categorical ambiguity resolution in this sub-section. This model will assume that a speaker starts his/her task of ambiguity resolution with lexical knowledge alone. Post-lexical, including structural and contextual, information will then be incorporated to improve on the result of lexical resolution. Thus, we can estimate baseline performance of categorical ambiguity resolution by using the lexical default frequency data. As mentioned earlier, lexical categorical assignment and ambiguity resolution is successful in two scenarios: where there is no ambiguity and where ambiguity goes to the default category. The baseline performance is very important in two fields of language processing studies: In computational linguistics, it represents the benchmark that any automatic tagging/ambiguity resolution program must surpass. In psycholinguistics, it represents he benchmark for valid subject performance as well as for evaluating model simulation.

A rough estimate of the performance can be given before our computational simulation. As mentioned above, 54.59% of the tokens in the corpus are categorically ambiguous, and the average default precision for all ambiguous tokens is 88.37%. Based on these two numbers, and assuming that

there is no unknown word, we estimate the precision rate of lexical category assignment at 93.65% (100%-54.59%+(54.59%x88.37%)). This is actually higher than most of the results reported by automatic tagging programs.

To simulate human ambiguity resolution, however, it is not adequate to simply obtain the statistics from a complete corpus. In actual language use, a speaker's task is to successfully process a text. A text could be a natural segment of a running dialogue or a natural segment of a written document. To simulate actual linguistic tasks, we take marked and segmented texts from Sinica Corpus as natural units of ambiguity resolution and category assignment. Since we are testing the simple lexical resolution and assignment algorithm, there is no speaker variation except in individual mental lexicons. Since we do not have individual mental lexicons, we will simply assume that a complete lexicon compiled from the corpus is a shared passive lexicon of all speakers. In other words, the only variation lies in the texts, and the experiment will show how well lexical resolution of categorical ambiguity resolution will perform assuming a perfect lexicon. We estimate the expected performance based on the average of all texts. In this experiment, we randomly picked one hundred and twenty (120) texts from Sinica Corpus. The length of texts ranges roughly from 900 to 90,000 characters (roughly from 500 to 50,000 words). These texts represent different topics, genres, styles, etc. We assume that there are no unknown words. Some of the typical results as well as the average results of the 120 texts are given below in Table 4.

**(4) Experiment Baseline Performance of Default Category Assignment**

| File name | Ambiguity | Default precision | Baseline res. Accu. |
|---|---|---|---|
| Nroot.tag | 54.7556% | 79.8791% | 88.9788% |
| Ointer.tag | 61.8716% | 92.3125% | 95.2436% |
| Okuo2b.tag | 59.9797% | 84.2452% | 90.5661% |
| Opinyi1.tag | 67.8163% | 87.3529% | 91.4232% |
| Opinyi2.tag | 64.8642% | 84.5382% | 89.9708% |
| Tsa.tag | 43.2161% | 91.2791% | 96.2312% |
| Ywcomp.tag | 52.7599% | 93.0657% | 96.3415% |
| Ywsubs.tag | 45.6564% | 91.2732% | 96.0157% |
| AVERAGE | 54.4011% | 89.0696% | 94.0537% |
| Corpus | 54.59% | 88.37% | 93.65% |

The lowest and highest number from this experiment are highlighted with underline to show the range of variation. The average mean of testing 120 texts are given right above the result obtained from the complete corpus (as if it were a single text) for contrast. The results show that even though degree of ambiguity (based on token frequency) in a text can vary, from as low as 43.21% to as high as 67.82%, it is somehow compensated by the changes in the frequency and dominance of default categories. Similarly, frequency of default category also varies widely, from a little under 80% to over 93%. There is a clear tendency for more ambiguous texts to also show high frequency of default category. The result is a fairly stable range of baseline ambiguity resolution

performance that averages at 94.05% but ranges, roughly, only from a little under 90% to a little over 96%.

The variation in degree of ambiguity as well as in dominance of default category is expected, as they are obviously dependent on the length, topics, etc. of each text. The stable baseline performance is surprising given the above variations. However, the surprising results itself lends strong support for a ambiguity resolution model based on lexical default. A human language cognition model would expect predictably good result of such basic linguistic task as category assignment, since a model that does not have good performance would predict frequent linguistic failure. And a model that allows too much fluctuation would predict that linguistic ability is highly context-dependent. Neither is correct. The fact that lexical default category assignment predicts the desirable result of categorical ambiguity resolution strongly suggest that it is the right model.

## 5. Categorical Sets and Categorical Ambiguity

In a random stochastic system, the distribution of samples is clearly dependent on the number of classes involved. This leads to an interesting question regarding categorization in human languages. Intuitively, there seems to be a 'universal' set of categories such that different languages can be compared with the same category set. And an identical grammar can be constructed by all speakers of the same language. However, in practice, we observe that different linguists often adopt different category sets that may vary widely in its size, from a score to a hundred. In addition, theoretical linguists often find it necessary to refer to the notion of sub-categories in accounting for linguistic facts. Thus, before asking the crucial question of if there is any cognitive or conceptual foundation of linguistic categorization, we need to ask if there is an optimal size of sets of categories.

The data involving categorical ambiguity offers the strictest test for the hypothesis that there be an optimal number of categories in human linguistic categorization. Again, in a simplistic formal system assuming no conceptual dependencies among categories, the probability for a given category to shift to any other category is the same. Hence degree of ambiguity will most likely be dependent upon the number of possible shifts, i.e. the number of categories. Furthermore, this model predicts that the number-of-classes effect on degree of ambiguity will show up as long as there is no complete dependency between two category sets of different sizes. In other words, if no sets of categories are optimal, then their size should play a role in the degree of categorical ambiguity. Thus, the number-of-classes effect will disappear only when one category set is optimal (or is a logically necessary reduction towards the optimal set by the larger set).

Sinica Corpus can be interpreted with two different category sets of either 13 or 44 categories, as introduced earlier. Hence it offers the opportunity to compare degree of ambiguity with different category sets as well as testing the hypothesis that there be an optimal number for the size of linguistic categories. In the following diagram, two frequency-ambiguity correspondence lines based

on the 44-category and 13-category tagging are put together for comparison. Note that this diagram is slightly different from diagram V since a smaller frequency range of 100 is taken, and hence the diagram is less smooth, though the positive dependency remains identical.

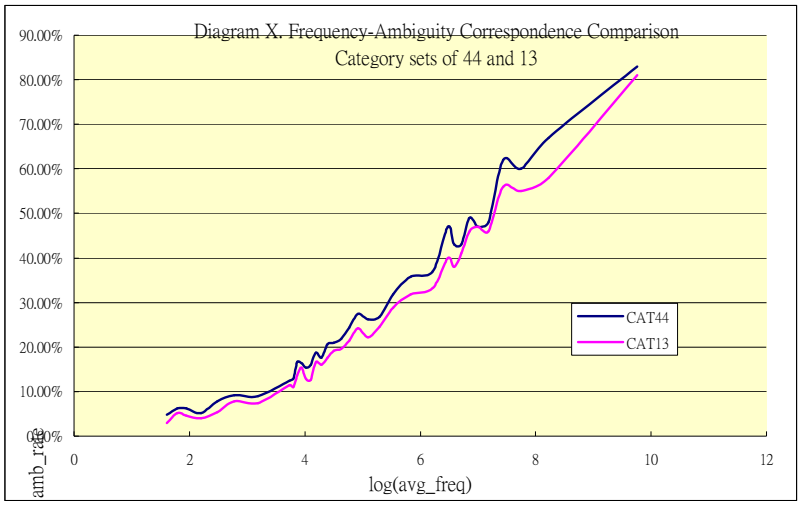



Diagram X, as well as all the other diagrams involving two different category sets: diagrams V, VI, VII, VIII, and IX, tell the same story. It contradicts naïve static intuition to find that the correspondence is essentially the same regardless of whether the data is taken from a tagsets of 44 or 13 categories.

As mentioned earlier, the 13 category set subsumes the 44 category set. In other words, the 44 category set can be regarded as the elaboration of the 13 category sets with sub-categories. The most plausible explanation for this contradiction to normal stochastic distribution is that the 13 categories represent the conceptually primary, hence optimal, categories. Hence all bona fide categorical ambiguity exist among these categories, whereas the added categories in the 44 categories sets are sub-categories that are mostly motivated by distributional difference. Since distributional differences are complementary by nature, there are few, if any, additional instances of categorical ambiguity among these sub-categories themselves.

To sum up, our study involving two different category sizes lends strong support to the conceptual primacy and psychological felicity of the small (13) category set. We find that this set is likely to represent the optimal size of category set (for Chinese). This finding also strongly suggests that categorical ambiguity is conceptually based, since pure structural categorization (i.e. subcategories added in the set of 44) surprisingly failed to increase categorically ambiguity. Thus it lends further support for future study to look for

conceptual/cognitive foundation of categorical ambiguity.

## 6. Future Studies: the 'Hard' Problems in Categorical Ambiguity Resolution

In section 3 above, we showed that there is a strong tendency towards a default category for categorically ambiguous items. In this penultimate section, we would like to find out if there are lexical items where such defaults do not exist. In other words, are there lexical items where ambiguity resolution must be dependent primarily on contextual information. We will refer to these items as ***Categorically Ambivalent***. The essential distributional data involving potentially ambivalent words are given below in Table 5.

**(5) Statistics of Ambiguous Words with Low Default Category**

| $P_{default}$ | No. of Words | % by Type | % by Token | % in Corpus |
|---|---|---|---|---|
| 0.51 | 132 | 3.14 | 1.95 | 1.06 |
| 0.52 | 164 | 3.90 | 2.09 | 1.14 |
| 0.53 | 210 | 4.99 | 2.31 | 1.26 |
| 0.54 | 254 | 6.04 | 2.64 | 1.44 |
| 0.55 | 299 | 7.11 | 2.80 | 1.53 |
| 0.56 | 335 | 7.96 | 3.08 | 1.68 |
| 0.57 | 355 | 8.44 | 3.16 | 1.72 |
| 0.58 | 399 | 9.48 | 3.41 | 1.86 |
| 0.59 | 448 | 10.65 | 3.69 | 2.01 |
| 0.60 | 506 | 12.02 | 4.75 | 2.59 |

Total Number of Ambiguous Words with Frequency Over 10 = 4,208

The statistics given above shows that only a small portion of all lexical items are categorically ambiguous. And among this small set, only a small portion is ambivalent. Since the effectiveness of lexical default model lies in the clear advantage it offers for choosing the default category, the potentially problematic cases are when choosing the default offers no clear advantage. For instance, if a default category has a frequency of 55%, then it could be only 10% stronger than a competing category. It its frequency is 60%, it can be only 20% stronger than a competing category. If 55% is taken as the cutoff point, then there are only 299 words that are ambivalent. In addition, they take up only 2.80% off all ambiguous words, and 1.53% of all corpus. The small number of ambivalent words as well as there limited representation in actual use suggest a lexicon-driven model for category ambiguity is still possible. In other words, even if each ambivalent word requires a different set of heuristic rules to resolve its ambiguity, they can still enumerated and be treated as lexical idiosyncracies.

In future studies, we will concentrate on these ambivalent words to find out if there is any cognitive motivation for the lack of a strong lexical default, as well as if such cognitive motivation can render ambiguity resolution cost-effective in context. Three representative examples are given below to suggest the kind of conceptual dependencies that may be involved.

**(5) Categorically Ambivalent Words**

| | | | |
|---|---|---|---|
| a. 像 *xiang4* | freq.=4414 | 1. | prep. 'like' 2342 |
| | | 2. | verb 'to be like' 2022 |
| | | 3. | noun 'likeness(=image/statue)' 50 |
| b. 影響 *yin3xiang3* | freq.=2397 | 1. | noun '(the) influence' 1291 |
| | | 2. | verb 'to influence' 1106 |
| c. 自然 *zi4ran2* | freq.=2164 | 1. | noun 'nature' 1012 |
| | | 2. | adverb 'naturally' 781 |
| | | 3. | adjective 'to be natural' 371 |

## 7. Conclusion

In conclusion, we have shown that lexical knowledge is crucial in categorical ambiguity representation and resolution. We also showed ambiguity is dependent on frequency and categorical identity. On the other hand, we showed that degree of ambiguity is not dependent on the size of categorical sets. Based on the fact that increase in number of categories does not noticeably increase degree of ambiguity, we argue that there is an optimal size of grammatical categories, probably around 13. Based on the fact that distributional sub-categories do not contribute to categorical ambiguity and that verbs are more likely to be categorically ambiguous than nouns, we suggest that categorically ambiguity might be primarily motivated by conceptual necessity. In terms of processing, we also established the baseline performance for categorical ambiguity resolution for Chinese, as well as identify the small number of lexical items that are hard problems for ambiguity resolution.

There are three important direction for future studies to take: First, we need to develop a default inheritance model for lexical representation that can both felicitously account for categorical ambiguity as well as offer an efficient algorithm for ambiguity resolution. Second, we need to find explanatory accounts for categorically ambivalent words, and hope that generalizations can be reached from these accounts to give them conceptual and cognitive motivation. Last, based on the above results, we need to explore the relationship between categorical ambiguity and semantic ambiguity. We suspect that conceptual basis can be found for both kinds of ambiguity, with the difference being that categorical ambiguity involves highly grammaticalized concepts, i.e. the linguistic categories.

## References


Abeille, Anne, L. Celement, and R. Reyes. 1998. TALANA Annotated Corpus: the first results. Proceedings of First Conference on Linguistic Resources. 992-999. Grenade, Spain.

Ahrens, Kathleen. 1998. Lexical Ambiguity Resolution: Languages, Tasks and Timing. In D. Hillert Ed. Sentence Processing: A Cross-linguistic Perspective. 11-31. New York: Academic Press.

Ahrens, Kathleen. 1999. The Mutability of Noun and Verb Meaning. Yuen-mei Yin, I-li Yand and Hui-chen Chan Eds. Chinese Languages and Linguistics V: Interactions in Language. 335-371. Taipei: Institute of Linguistics, Academia Sinica.

Briscoe, Ted., V.d.Paiva, and A. Copestake. eds., 1994. Inheritance, Defaults, and the Lexicon. Cambridge: Cambridge University Press.

Cavalli-Sforza, Luigi L. 1994. An Evolutionary View in Linguistics. In M. Y. Chen and O. J-L. Tzeng. Eds. In Honor of William S-Y. Wang. Interdisciplinary Studies on Language and Language Change. 17-28. Taipei: Pyramid.

Cavalli-Sforza, Luigi L, and William S-Y. Wang. 1986. Spatial Distance and Lexical Replacement. Language. 62.38-55.

Chao, Yuen Ren. A Grammar of Spoken Chinese. Berkeley: University of California Press.

Chen, Keh-jiann, Chu-Ren Huang, Li-ping Chang, and Hui-Li Hsu. 1996. Sinica Corpus: Design Methodology for Balanced Corpora. In. B.-S. Park and J.B. Kim. Eds. Proceeding of the 11th Pacific Asia Conference on Language, Information and Computation. 167-176. Seoul:Kyung Hee University.

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 Pacific Asia Conference on Formal and Computational Linguistics. 81-95. Taipei: R.O.C. Computational Linguistics Society.

Chinese Knowledge Information Processing (CKIP) Group. 1995. An Introduction to Academia Sinica Balanced Corpus for Modern Mandarin Chinese. CKIP Technical Report. 95-01. Nankang: Academia Sinica.

Chomsky, Noam. 1995. The Minimalist Program. Cambridge: MIT Press.

Gentner, Dedre. 1981. Some Interesting Differences Between Verbs and Nouns. Cognition and Brain Theory. 4:2.161-178.

Gentner, Dedre. and Ilene France. 1988. The Verb Mutability Effect: Studies of the Combinatorial Semantics of Nouns and Verbs. In Small et al. 1988.343-382.

Gorfein, D.S. 1989. ed., Resolving Semantic Ambiguity. New York: Springer Verlag.

Huang, Chu-Ren. 1999. *SouWenJieZi 搜文解字:*A Linguistic KnowledgeBase Anchoring Chinese Digital Museums. Presented at Digital Museum Seminar and AP Digital Library Consortium Joint Meeting 1999. July 21-23. Taipei.

Huang, Chu-Ren. 1994. Corpus-based Studies of Mandarin Chinese: Foundational Issues and Preliminary Results. In M. Y. Chen and O. J-L. Tzeng. Eds. In Honor of William S-Y. Wang. Interdisciplinary Studies on Language and Language Change. 165-186. Taipei: Pyramid.

Huang, Chu-Ren, Zhao Ming Gao, Claude C.-C. Shen, and Keh-jiann Chen. 1998a. Quantitative Criteria for Computational Chinese Lexicography: A study based on a standard reference lexicon for Chinese NLP. Proceedings of ROCLING XI.87-108.

Huang, Chu-Ren, Keh-jian Chen, Zhao Ming Gao, Feng Yi Chen, and Claude C.-C. Shen. 1998b. Word Frequency Dictionary. CKIP Technical Report 98-01. Nankang, Taipei: Academia Sinica.

Huang, Chu-Ren, Keh-jian Chen, Zhao Ming Gao, FengYi Chen, and Claude C.-C. Shen. 1998c. Accumulated Word Frequency in Sinica Corpus. CKIP Technical Report 98-02. Nankang, Taipei: Academia Sinica.

Huang, Shuanfan. 1994. Chinese as a Metonymic Language. In Mathew Y. Chen and Ovid J.-L.. Tzeng. eds., In Honor of William S-Y. Wang. Interdisciplinary Studies on Language and Language Change. 223-252. Taipei: Pyramid.

Lyons, John. 1977. Semantics. Cambridge: Cambridge University Press.

Manning Christopher D. and Hinrich Shutze. 1999. Foundations of Statistical Natural Language Processing. Cambridge: MIT Press.

Meng, Helen and Chun Wah Ip. 1999. An Analytical Study of Transformational Tagging for Chinese Text. In Proceedings of ROCLING XII. 101-122. Taipei: Association of Computational Linguistics and Chinese Language Processing.

Pustejovsky, James. 1995. The Generative Lexicon. Cambridge: MIT Press.

Redington, Martin, Nick Chater, Chu-Ren Huang, Li-ping Chang, Steve Finch, and Keh-jiann Chen. 1995. The Universality of Simple Distributional Methods: Identifying syntactic categories in Mandarin Chinese. Paper presented at the International Conference on Cognitive Science and Natural Language Processing. July 7-11. Dublin City University.

Schutz, Hinrich. 1997. Ambiguity Resolution in Language Learning: Computational and Cognitive Models. Stanford: CSLI Publications.

Small, S.I., W.C. Garrison, and M.K. Tanehaus.1988. (Eds.) Lexical Ambiguity Resolution. San Mateo: Morgan Kaufmann.

Wang, William S-Y. 1991. Explorations in Language. Taipei: Pyramid.

Wang, William S-Y. 1969. Competing Changes as a Cause of Residue. Language. 45.1.9-25.