Noun Class Extraction from a Corpus-based Collocation Dictionary: An Integration of Computational and Qualitative Approaches

HUANG Chu-Ren*, CHEN Keh-Jiann* & GAO Zhao-Ming#

*Academia Sinica, #Ji-Nan International University

Abstract

Classifiers are widely believed to categorize noun classes (e.g. Tai 1994). However, the system of classification as well as its explanation has been subject to different interpretations. The exact nature of the classifier system is even more intriguing when classifiers encode several semantic dimensions; hence more than one classifier may co-occur with a noun, often restricting or coercing the nominal meaning (Ahrens and Huang 1996). In this paper, we integrate computational and qualitative approaches to (linguistic) knowledge acquisition and manipulation.

1. Background

The main question we want to answer is whether there is a way to objectively and effectively capture human being's linguistic knowledge? The traditional intuition-based approach has often been criticized for being biased and failing to capture the fact that language and its grammar is the result of the collective behaviors of native speakers (e.g. Huang 1994). On the other hand, corpus-based approaches driven by statistics have been restricted by the scope of the data and the available methodology.

We have argued in Huang (1994), in agreement with many linguists who adopt corpus-based approaches, that the stochastic method based on large quantity of data does offer us insights not available with introspective observation. We will argue in this paper that if data-rich resources, such as corpora, and knowledge–rich resources, such as dictionaries and grammars, are integrated, then quantitative and qualitative approached to linguistics can be combined to offer interesting insights and generalizations.

In our proposal, the combination of the quantitative and qualitative approaches involves three crucial steps. First, one must use objective data that is part of the linguistic system. The noun-classifier collocation data is what is used in this study. The reason for choosing such data is two-fold: Firstly, the use of objective data avoids subjective judgements involving the speaker; secondly, data that describes a grammatical system not only offers direct evidence of grammatical representation but also avoids the difficulties entailed by the often incomplete or fragmental nature of corpus data.

Second, stochastic methodology is used to extract generalizations from the data. In our case, Shannon's information theory is adopted. A stochastic approach to linguistic generalizations is able to take account of large quantity of data as well as avoid the danger of human biases.

Third, the interpretation and explanation of the generalizations will be carried out by linguists. The quantitative approach can only carry us as far as the data goes. An explanatory theory must be formulated through the analysis of quantified data and generalizations, and synthesis of the analysis with model-theoretical considerations.

1.1. The Data: First Integration of Corpus Collocation Data and Human Interpretation

The data that will be used in this study comes from the Mandarin Daily Classifier Dictionary (Huang et al. 1997). The dictionary itself is the result of our first attempt to integrate corpus collocation data with human analysis. Our approach, as reported in Chang et al. (1996) consists of the following steps:

1. Collection of 1) classifiers, 2) nouns (according to the identity of their stem/final), and 3) noun-classifier co-occurrences directly from the Sinica Corpus
2. Sort the above data according to frequency
3. Linguists/lexicographers make generalizations based on the above sorted data

The result is a collocational dictionary of nouns, exemplified by diagram 1. The nouns are listed according to their head. Under each entry sharing the same head noun, sub-entries are listed according to the natural classes formed by nouns sharing the same set of collocating classifiers. We hypothesize that each sub-entry is a sense. This hypothesis is reinforced by the fact that it is a straightforward task for linguist/lexicographer to provide notes to define each sense. In other words, these data are quantifiable qualitative characterizations of the grammar. They are quantifiable because the number and information content of the collocating classifiers can be calculated. The information-theoretic foundation of such calculation will be laid in the next section.

1.2. Theoretical Foundation: Language as an Information System

The crucial theoretical question that we will be asking in this paper is the following:

> Can the Saussurean definition of grammar as a structured system of SIGNs be reinterpreted as a **structured system of codes+information**?

There have been a few linguistic studies that reinterpret Saussure's sign as a bearer of information, including HPSG (Pollard and Sag 1987 and 1994), and ICG (Chen and Huang 1990 and 1996). Hence, it is natural to reinterpret the signifier/signified contrast as the code/information contrast in Shannon's (1949) Information Theory. This re-interpretation should provide a formal framework to quantify the linguistic information involved. However, it remains unclear how relevant linguistic information should be quantified. The critical issue is: when Shannon's Information Theory is adopted in a linguistic study, is it the information content or coding complexity that is being quantified?

The theory of entropy is indeed a theory about the complexity of the coding system. However, it is reasonable to assume that the complexity of the coding must be motivated by the richness of the information content. Even though most linguists follow Saussure and maintain that the choice of sign is arbitrary, this position does not entail that the choice of the signing structure is also arbitrary. Since our concern is the substantive explanation of linguistic facts, we would like to find a positive answer to the following question.

> Can a quantitative measure (of information load) lead to generalizations concerning information content?

1.3. The Role of Corpus in Quantitative and Qualitative Studies

The essential data that a corpus offers is the distributional information of linguistic elements. However, distribution is NOT a good clue for determining conceptual primes (Huang et al. 1998a,b). For example, Zipf's Law predicts the normal distribution of lexical entries in actual use. Huang et al. (1998 a. b) reports that the distribution of a set of proposed semantic primes in the Sinica Corpus follows Zipf's Law. This means that semantic primes do not differ from other lexical items in terms of distribution; moreover, there are no distributional cues found to identify them.

However, on the other hand, corpus when used as primary data does offer more comprehensive coverage, as well as more comprehensive evidence for a grammatical system (Chang et al. 1996). In terms of Information Theory, information load and content cannot be studied unless the whole signing system can be described. The distributional nature and the comprehensive representation of a corpus offers a good basis for describing a complete linguistic sub-system, such as the noun-classifier system. In other words, even though the statistics of distribution itself cannot lead to direct linguistic characterization, the actual set of collocation data still provides a solid basis for our description and representation of the linguistic sub-system. This representation is crucial for our qualitative and quantitative characterization of the information content.

1.4. Our Methodology

First, since the collocation of a classifier and a noun signifies the categorization of nouns, we exhaustively list the collocational relationship (Chang et al. 1996). Second, based on the collocational information (Huang et al. 1997), we calculate the information load of each classifier when it co-occurs with noun. Third, based on the information load of all collocating classifiers, we calculate the affinity of noun classes to establish the noun class system as encoded by Chinese classifiers.

We adopt Shannon's information theory (Shannon and Weaver 1949, Pierce 1980) to quantify the information load of classifier-noun collocation. The information load of any classifier-noun collocation is the entropy difference between all noun classes and all noun classes that co-occur with a particular classifier. The categorization information of a noun is a vector recording the information load of the collocating classifiers (i.e. a particular position will have the value 0 if the classifier does not co-occur, or a value calculated above if it does). Information distance among these noun classes will be calculated based on these vectors.

2. The Grammar of Mandarin Chinese Classifiers

The classifier system is an important characteristic of Chinese grammar, hence it has received much attention in the literature. Based on our survey of the literature, we draw our linguistic account mainly from the following papers: Ahrens and Huang (1996), Huang et al. (1997), and Tai (1994).

Tai (1994), among others, noted the crucial distinction between *bona fide* classifiers and measure words. Measure words are standardized units used to measure certain physical properties of an object, such as weight, height, age etc. In other words, they measure the common physical properties but do not refer to the properties that differentiate noun classes. Following this account, we do not include the linguistic data involving measure words in our current study. The four classes of classifiers that we will consider are Individual Classifiers, Mass Classifiers, Kind Classifiers, and Event Classifiers. This classification follows Huang et al. (1997).

Individual Classifiers, such as 一個人 one-CLS-person 'a person', are the prototypical classifiers that refer to certain properties of individuated entities and thus help to identify different noun classes. The other three types of classifiers also define noun classes, but they do to refer to classical individuals. First, Mass Classifiers classify Mass nouns. When used with an individual-denoting noun, it coerces a Mass reading 一群人 one-CLS-person 'a group of people'. Second, Kind Classifiers differentiate different *kinds* of nouns, such as 一款車 one-CLS-car 'this design of car, referring to cars having the same design'. Last, Event Classifiers classifier different event nominals, whether base or derived. For instance, 一通電話 one-CLS-phone 'a phone call, referring to the calling event', as opposed to 一支電話 one-CLS-phone 'a telephone (machine)'. How classifiers coerce collocating nouns into different types is discussed in Ahrens and Huang (1996).

Two facts of the grammar of classifier-noun collocation in Mandarin are crucial in our current study: First, noun classes are not defined by any single classifier. On one hand, each classifier refers to a particular semantic/conceptual property that may be shared by many classes of nouns. On the other hand, a noun does not select a single classifier. It allows the collocation of any classifier in a context where they refer to the same semantic properties. Hence, it is the combination of all the collocating classifiers that more fully describes the semantic content of a noun class. Thus, the information content of a noun class must include all the information carried by all the classifiers that can collocate with this noun class. To account for this fact, we propose to represent the information content in a vector model, where the nominal semantic information is represented by a vector determined by the composition of individual components representing each collocating classifier.

The second fact is that conceptual/semantic classifications are logical ones that are not subject to fuzzy interpretations influenced by frequency. In other words, the frequency of use of a particular sense of a noun has nothing to do with whether it entails a particular semantic property or not. This is another point supporting the position that distribution statistics is not directly useful in semantic studies. To make sure that all collocational information from classifiers are accounted for, we use the knowledge-rich Noun-CLS collocation dictionary (itself based on rich data from corpus). In other words, we are taking into account the whole grammatical system of nominal semantic classification knowledge.

Two caveats need to be pointed out before we go into how the classifier information is used to extract noun classes. First, it is crucial to note that not all classifiers have the same classificatory power. For instance, 種 zhong3 and 個 ge1 co-occur with most noun classes and have very low differentiating power. This fact is important when we want to decide whether one noun class is semantically closer to another one. Thus we need a reliable measure for the information load of classifiers. Second, it should noted that the non-collocation of a certain classifier also carries information. If a classifier cannot collocate with a noun, it indicates that the semantic properties denoted by that classifier is either not part of the semantics of that particular noun or is incompatible with it. Ideally, such information should also be utilized when we automatically divide noun classes. The two above observations leave us in a dilemma regarding how to deal with non-collocating classifiers. In short, we are able to estimate the classificatory power of a classifier when it co-occurs with nouns, but not able to quantify its significance when it does not occur. Thus we adopt the null hypothesis that non-collocation does not carry significant information and assign the value of 0 to it. We explain how to calculate the information load for collocating classifiers in the next section.

3. The use of Information: Entropy and Information load based on entropy

3.1. A Quantifiable Definition of Information Load

Shannon's definition of **Entropy** is actually a measurement of information content. In an *i*-item system where p_i is the assigned probability of each item, the Entropy of the system is defined as follows:

(1) Entropy = $- \Sigma p_i \log_2 (p_i)$

One of the premise of this paper is that classifiers carry noun-classification information. Thus, the information content that concerns us here is the semantic/conceptual information. However, we now know that distributional information such as frequency does not reflect semantic information, and that there is no other objective measure for the event probability of the occurrences of semantic properties. Thus we start with the null hypothesis that each event involving a noun group with a classifier collocation has equal probability. If we assume equal probabilities for each item/sign, then the entropy of an N-sign system can be calculated as:

(2) $- \Sigma 1/N \log_2 (1/N) = -\log_2 (1/N) = \log_2 N$

Thus, the entropy (i.e. information content) of a classifier X is equal to the entropy difference between the whole nominal system and the sub-system that collocates with it. Again, since we assume that each noun class represents a natural semantic/conceptual group, its frequency is irrelevant to the semantic interpretation, and we can assume equal probability for the event of the occurrence of each noun class. In other words, if a classifier X collocates with n of the N possible classes of noun, then its contribution to the information of the system can be calculated as:

(3) $\mathrm{Log}_2 (N) - \log_2 (n)$

Under this interpretation, with N being constant, it is the classifiers that collocate with the least number of noun classes that bear the most information content.

3.2. The Clustering Algorithm

The above algorithm allows us to represent the information load as well as describe the topology of the sub-grammar of classifier-noun collocation. However, we should reiterate that we still have not obtained any direct description of the information content. However, with the above quantitative description, we are able to measure the distance between any two noun classes in the multi-dimensional space defined by the classifiers. Since the information load of the classifier-noun collocation system is semantics-based, we assume that the distance also reflects the semantic affinity between these two noun classes. Two test this hypothesis and to explore the possibility of the combination of quantitative and qualitative studies of linguistics, we form a semantic tree of noun classes by iterativelly joining the two most similar noun classes based on their classifier collocation information. Our algorithm for clustering noun classes follows:

1. Calculate the information load of the 182 classifiers using the above definition (3)
2. Each classifier corresponds to a dimension in the vector space of 182 dimensions. The projection length of each dimension is the information load of each classifier, as defined above.
3. Each group of nouns is assigned a vector defined by the summation of all vector dimensions corresponding to classifiers that collocates with this group.
4. Iteratively cluster any two groups of nouns with minimal distance to create a new noun group and assign it a vector of equal distance to the two original vectors (i.e. V1+V2/2). This step is carried out continuously until there is only one single one noun class left.

Step 1 calculates the information load of each classifier. The result shows that the classifier with the lowest information load (i.e. entropy) of 1.269 is 個 ge5, the well-known general and neutral classifier. And the second lowest, with entropy of 3.363, is 名 ming2, the general classifier for humans. These are followed closely by 位 wei4, 一點 yi1dian3, 群 qyun2, and 隻 zhi1; all general classifiers that occur with a large range of nouns. The classifiers that have the highest information load of 11.52 in our system are those which uniquely identify a noun group, such as 闋 que4, 題 ti2, 桌 zuo1, and 班 ban1.

Take note that in step 3 above, dimensions of non-collocating classifiers are not included. The vector representing each noun group is defined by all classifiers collocating with this noun group, corresponding to the linguistic account that all collocating classifiers jointly describes the semantic properties of the noun group. Hence, noun groups that share an identical set of collocating classifiers cannot be differentiated and are lumped as one class and assigned a unique vector. Thus even though there are 1,910 nominal endings and over 2,000 entries (including idiosyncric nouns that do not share an ending with others) in the collocational dictionary, the noun-classifier collocation system only identifies 502 different noun groups that can be assigned vector values.

The result of the above algorithm is a binary tree with a single mother node. The terminal nodes are the noun classes as defined by classifier collocation. Any noun class combines with another class that is closest to it and form a new class. The procedure is conducted iteratively until all noun classes are joined under one mother node. We attempt to interpret all intermediate nodes.

Whether our assumptions and this above algorithm is credible depends on whether the tree constructed can be given a reasonable semantic account. In other words: does this algorithm produce a possible semantic network for Chinese nouns? If the answer is positive, then we not only prove that Chinese classifier system is semantically-based, but also that corpus-based quantitative methods are more versatile than one might think. On the other hand, if the clustering result is unsuccessful, we will need to think more about our premise and formal assumptions.

4. Interpretation of the result:

Our preliminary examination shows that reliable classification results are obtained for sub-trees with depth of less than 4. Groupings are largely counter-intuitive for tree depth of 5 or deeper. In other words: we have obtained somewhere between 50 to 75 valid and interesting noun classes by this method. Two of such noun classes are given in Diagrams 2a and b.

We think that the possible reasons for the failure of taller trees are the following: First, the medium distance approach of vector combination may be incorrect. It allows some factors to be diluted too fast. In other words, because no information load is assigned to no-collocating classifiers, we are not able to make the distinction between semantic contradiction and semantic irrelevance. Theoretically, two sub-classes with contradictory semantic properties should cancel each other out. This means that this particular property is irrelevant in characterizing the combined class. However, if the non-occurrence is only due to irrelevance for that sub-class, then a semantic property marked on the other sub-class should still be inherited by the combined class, albeit with less characterizing power. To solve this problem, a more sophisticated model needs to be devised. However, such a model would also require that there is comprehensive dictionary data marking the semantic causes of all non-collocating noun-classifier pairs. Methodologically, this will involve controversial speaker judgement that cannot be verified with empirical data. Thus we will not pursue this line of thinking further.

Second, take note that classifiers are often ambiguous themselves. For instance, the classifier 條 tiao2 can indicate the semantic properties of 1. narrow and long objects, 2 slender animals, 3, narrow channel/conduit, 4, the semantic class of 'line', 4, the semantic class of 'law, rule', 5, the semantic class of 'life', and 6, the semantic class of 'song'. This current study takes each classifier as a unique sign and does not differentiate its different semantic properties. This has the advantage of processing convenience but incorrectly group different semantic properties to the same class simply because they share a sign with the same form. We are currently using information from both the classifier dictionary and the noun-classifier collocation dictionary to obtain a finer-grain collocation relation that take into account all the senses of each classifier. (In fact, work carried out to date in this vein shows that the initial number of noun groups that can be identified more than doubles, from 502 to 1,061.) As a result, many noun groups that are semantically different but incorrectly grouped together because of the coarser granularity of the current approach can be correctly separated. This should lead to improved results in noun clustering.

5. Conclusion

Nominals bear a crucial information load in communication yet their semantic structures are often difficult to determine because of the rather productive semantic process of type-shifting (Ahrens and Huang 1996). We show with Mandarin that certain types of semantic information can be explicitly marked by linguistic cues. Utilizing information such as the clustering of collocating classifiers and the grouping of nouns sharing the same head morpheme, we proposed an approach to automatically extract nominal semantic structures from corpus. This proposal is an example of how quantitative and qualitative approaches can be productively combined in linguistic studies.

由名量搭配辭典中自動抽取名詞語意分類架構：計質與計量方法之結合

黃居仁*， 陳克健*， 高照明#

*中央研究院，#暨南國際大學

摘要

學者們大致同意中文分類詞的主要功能是對名詞的語意作分類(Tai 1994)。但對於分類詞的分類系統分類的語意或觀念解釋，卻較缺乏一致的看法。本文結合計質與計量方法，利用直接導自語料庫的名詞量詞搭配辭典訊息，自動推導出中文名詞觀念及語意分類系統。我們採用單農（Shannon and Weaver 1949）的訊息理論（Information Theory）來計算名量搭配的訊息內容，並提出用向量(Vector)模式來計算名詞群間的距離。我們用以上的模式推導出一個大致可行的漢語名詞語意分類架構。

References

Ahrens, Kathleen, and Chu-Ren Huang. 1996. Classifiers and Semantic Type Coercion: Motivating a New Classification of Classifiers. Proceedings of PACLIC11. 1-10. Seoul: Kyung Hee University.

Chang, Lili, Keh-jiann Chen, and Chu-Ren Huang. 1996. The Use of Corpus in Dictionary Compilation.[In Chinese] Proceedings of ROCLING IX. 255-279.

Chen, Keh-jiann and Chu-Ren Huang. 1996. Information-based Case Grammar: A Unification-based Formalism for Parsing Chinese. In Huang et al. 1996. pp.23-46.

Chen, Keh-jiann and Chu-Ren Huang. 1990. Information-based Case Grammar. Proceedings of the 13th International Conference on Computational Linguistics (COLING). Vol.ii.54-59. Helsinki, Finland.

Huang, Chu-Ren. 1994. Corpus-based Study of Chinese: 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. 479-494. Taipei: Pyramid.

_____ 1998. Classified Information: A Corpus-based Approach Towards Automatic Extraction of Nominal Semantic Structures. Invited talk, the Third International Conference on Information-Theoretical Approaches to Language, Logic, and Computation (ITALLC3). Hsitou, Taiwan, June 16-19, 1998

Huang, Chu-Ren, Keh-jiann Chen, and Ching-hsiung Lai. 1997. Mandarin Daily Classifier Dictionary. Taipei: Mandarin Daily Press.

Huang, Chu-Ren Zhao-ming Gao, Keh-jiann Chen and Claude C.C. Shen. 1998a. Towards a Sharable and Reusable Lexical List: The construction of a standard reference lexicon for Chinese NLP. The Pacific Neighborhood Consortium Annual Meeting. June 16-18. Taipei: Academia Sinica.

Huang, Chu-Ren, Zhao-ming Gao, Claude C.C. Shen and Keh-jiann Chen. 1998b. Quantitative Criteria for Computational Chinese Lexicography. Proceedings of ROCLING XI. pp.87-108.

Huang, Chu-Ren, Keh-jiann Chen and Benjamin K. T'sou.1996. Eds. Readings in Chinese Natural Language Processing. Journal of Chinese Linguistics Monograph Series No. 9. Berkeley Journal of Chinese Linguistics.

Pierce, John R. 1980. An Introduction to Information Theory: Symbols, Signals, and Noise. Revised Edition. New York: Dover.

Pollard, Carl, and Ivan A. Sag. 1994. Head-Driven Phrase Structure Grammar. Stanford: CSLI. And Chicago: U. of Chicago Press.

Pollard, Carl, and Ivan A. Sag. 1987. Information-Based Syntax and Semantics. Volume I: Fundamentals. CSLI Lecture Notes. No. 13. Stanford: Center for the Study of Language and Information.

Shannon, Claude E. and W. Weaver. 1949. The Mathematical Theory of Communication. Urbana: University of Illinois Press.

Tai, James H.-Y. 1994. Chinese Classifier System and Human Categorization. In M.Y. Chen and O.J.-L. Tzeng Eds. In Honor of William S-Y. Wang: Interdisciplinary Studies on Language. Taipei: Pyramid.

Quantitative and Computational Studies on the Chinese Language
(Eds.) B.K. T'sou, T.B.Y. Lai, S.W.K. Chan & W. S-Y. Wang
《漢語計量與計算研究》
編輯: 鄒嘉彥、黎邦洋、陳偉光、王士元
©1998 Language Information Sciences Research Centre
City University of Hong Kong
ISBN: 962-442-135-8

Diagram 1.

Sample Entry of Noun-Classifier Collocation Dictionary (Huang et al. 1997): -fa3 'method, point (of view), law, rule, skill, power'

法ㄈㄚˇ

1. **方法或方式。**

◎ 方法或方式：方法、辦法、作法、做法、手法、用法、寫法、療法、玩法、演算法、…。〔一般〕：個、項、套、招、組。〔種類〕：樣、式。

◎ 意見：看法、說法、想法、講法、…。〔一般〕：個、項、點。〔種類〕：派、樣、式。

辨析：我們可以說「這一點看法、這一點說法、這一點想法」，但是不能說「這一點講法」。

2. **法律或規律。**

◎ 指各種法律：憲法、勞基法、刑法、民法、交易法、選罷法、國安法、著作權法、保育法、國際法、軍法、稅法、…。通常不搭配量詞

辨析：「憲法」還可以說「一部憲法」，指的是「憲法」這部書。法律條文的內容是依「條、項、款」編列，出現在法律名稱的後面，如「民法第一百八十條第一項第二款、公司法第四百一十九條第一項第五款」。

◎ 指語文的規律：語法、文法、句法、…。〔一般〕：套、條、個。

3. **技藝或法力。**

◎ 槍法、劍法、箭法、刀法、指法、…。〔一般〕：套、個。〔種類〕：派、式。

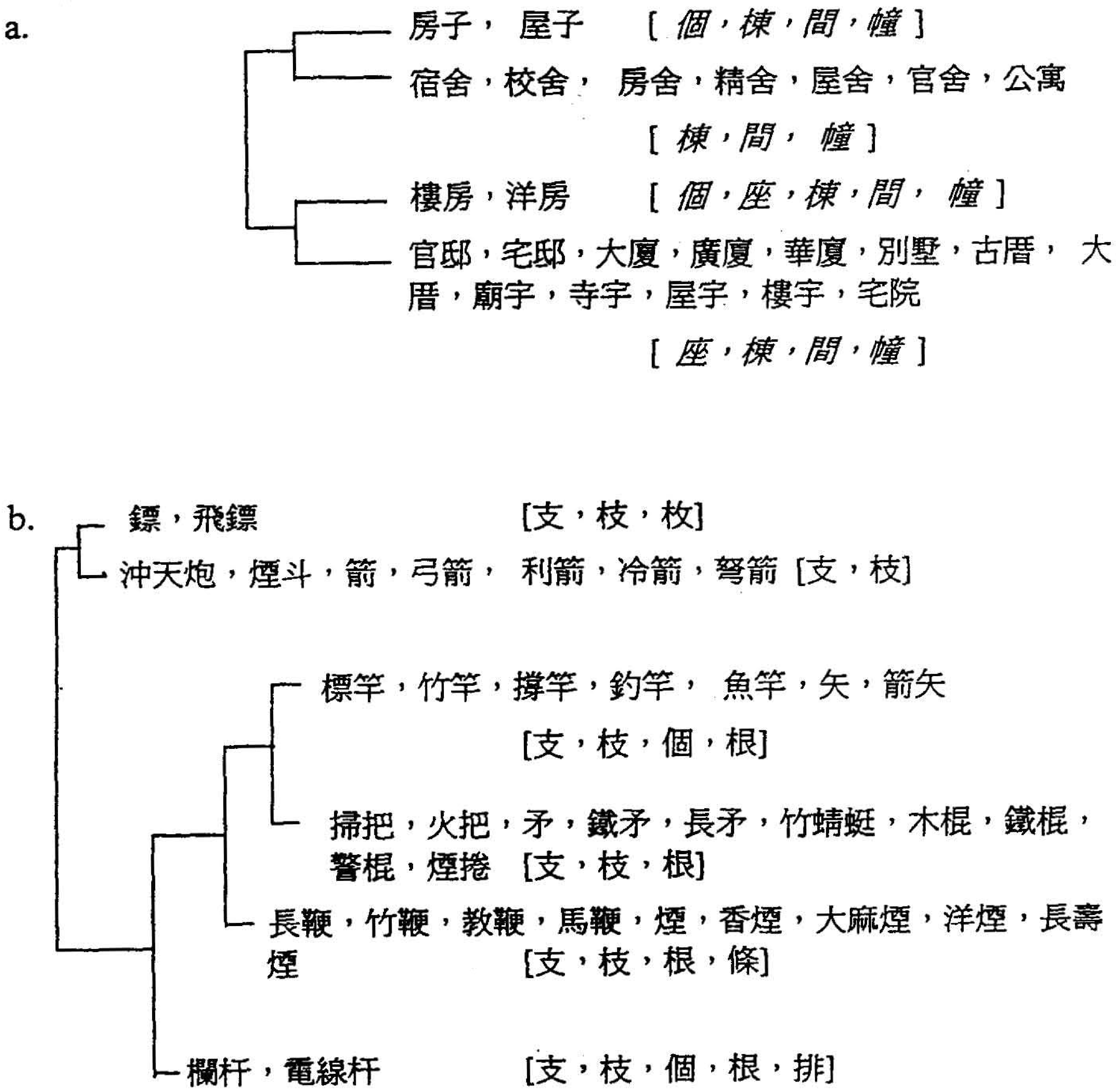
◎ 書法。〔一般〕：幅、張、篇、件。

辨析：「書法」除了和上述量詞搭配之外，還有「他寫得一手好書法」這樣的說法。

◎ 佛法、魔法、…。通常不搭配量詞。

Diagram 2.

Sample Sub-trees of Noun Cluster


a.
房子， 屋子 [個，棟，間，幢]
宿舍，校舍， 房舍，精舍，屋舍，官舍，公寓
[棟，間， 幢]
樓房，洋房 [個，座，棟，間， 幢]
官邸，宅邸，大廈，廣廈，華廈，別墅，古厝， 大厝，廟宇，寺宇，屋宇，樓宇，宅院
[座，棟，間，幢]
b.
鏢，飛鏢 [支，枝，枚]
沖天炮，煙斗，箭，弓箭， 利箭，冷箭，弩箭 [支，枝]
標竿，竹竿，撐竿，釣竿， 魚竿，矢，箭矢
[支，枝，個，根]
掃把，火把，矛，鐵矛，長矛，竹蜻蜓，木棍，鐵棍，警棍，煙捲 [支，枝，根]
長鞭，竹鞭，教鞭，馬鞭，煙，香煙，大麻煙，洋煙，長壽煙 [支，枝，根，條]
欄杆，電線杆 [支，枝，個，根，排]