A Preliminary Study of Disambiguating VO- and VN-Constructions Using Selection Preferences

Kok-Wee Gan
Department of Computer Science
Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong

E-mail: gankw@cs.ust.hk

Abstract

In Chinese, a verb followed by a noun can be analyzed as either a verb-object (VO) construction or a verb-noun (VN) construction. In the latter, the verb acts as a modifier of the noun. This paper describes how selection preferences can be used to determine whether a Verb+Noun pair (V+N) is a VO-construction or a VN-construction. The approach also takes syntactic factors into consideration. These factors are expresssed in terms of likelihood measures of the tendency of verbs and nouns functioning as VN- and VO-constructions. The preliminary result based on 17 bi-syllabic, transitive verbs with a total of 880 V+N pairs is 88.4%.

1 Introduction

In Chinese, a verb followed by a noun can be analyzed as either a VO-construction or a VN-construction. For example, 訓練口才 *xun4lian4 kou3cai2* 'train oratorical skills' is a VO-construction, where 口才 *kou3cai2* 'oratorical skills' is the object of the verb 訓練

xun4lian4 'train'. However, 訓練方法 *xun4lian4 fang1fa3* 'training methods' is a VN-construction, with the verb 訓練 *xun4lian4* 'training' acting as the modifier of the noun 方法 *fang1fa3* 'methods'. There is no inflections in Chinese to distinguish between these two usages of verbs. This ambiguity poses a problem for a Chinese parser. In this paper, we describe an approach to automatically determine whether a V+N pair is a VO- or VN-construction using selection preferences.

Selection preferences cast selection restrictions in probabilistic terms. Selection restrictions of a predicate are specifications of the necessary and sufficient condition for a semantically acceptable argument. Such conditions are represented as boolean functions of semantic markers. Selection preferences, in contrast, represent such conditions as real-value functions. Such conditions are usually derived from corpora. For example, semantically acceptable arguments which can be the object of the predicate 吃 *chi1* 'eat' tend to be *physical*, *animate*, *edible*, and so forth. Measures of how likely the object of 吃 *chi1* 'eat' is *physical*, *animate*, etc., constitute the selection preferences of 吃 *chi1* 'eat'. In Section 2, we describe an information-theoretic approach of determining the selection preferences of a predicate [7]. We will explain how we make use of selection preferences to disambiguate VN- and VO-constructions in Section 3. The experimental results will be reported in Section 4. A comparison with related work will be covered in Section 5.

2 Determination of Selection Preferences

We define the selection preferences of a predicate over a taxonomy of 116 conceptual classes [1]. The taxonomy is primarily organized into a hyponymy (IS-A) hierarchy as shown in the appendix. Some of the conceptual classes, for example, *edible, flowers, fruits, holes, human, literature,* and *locative,* are features that serve to link together concepts which are otherwise not related in the hierarchy. These concepts are listed in the appendix with a plus operator in front.

The information-theoretic approach of determining selection preferences as proposed in [7] is adopted and summarized as follows.

Let P be a random variable ranging over the set $\{p_1, p_2, ..., p_m\}$ of predicates. Let C be another random variable ranging over the set $\{c_1, c_2, ..., c_k\}$ of conceptual classes in a taxonomy. C is related to P by a particular predicate-argument relationship, such as verb-object, or adjective-noun. The preference of a particular predicate p_i is defined as the effect it has on the distribution of C. Let the distribution of C regardless of any particular predicate be the prior distribution, $\mathrm{p}(c)$, and let the posterior distribution $\mathrm{p}(c|p_i)$ be the distribution of C given the predicate p_i. The change between the prior distribution $\mathrm{p}(c)$ and the posterior distribution $\mathrm{p}(c|p_i)$ constitutes the selection preference strength of the predicate p_i, which can be measured by relative entropy. In information theory, the relative entropy of two probability distributions p and q is defined as:

$$\mathrm{D}(\mathrm{p}\|\mathrm{q}) = \sum_x \mathrm{p}(x)\log\frac{\mathrm{p}(x)}{\mathrm{q}(x)} \tag{1}$$

Replacing p with $\mathrm{p}(c|p_i)$, q with $\mathrm{p}(c)$, the **selection preference strength** of the predicate p_i is:

$$\mathrm{D}(\mathrm{p}(c|p_i)\|\mathrm{p}(c)) = \sum_c \mathrm{p}(c|p_i)\log\frac{\mathrm{p}(c|p_i)}{\mathrm{p}(c)} \tag{2}$$

The **selectional association** of a predicate, $\mathrm{A}(p_i, c_j)$, for a particular argument class c_j is defined as:

$$A(p_i, c_j) = \frac{1}{\eta_i} p(c_j|p_i) \log \frac{p(c_j|p_i)}{p(c_j)} \qquad (3)$$

where η_i is the selection preference strength of the predicate p_i as shown in equation (2).

The **selection preference** of a predicate p_i is a vector of selectional associations between p_i and a list of conceptual classes in a taxonomy. The statistical technique of maximum likelihood estimation (MLE) is used in deriving the prior distribution p(c) and the posterior distribution p($c|p_i$). For a particular conceptual class c_j, p(c_j) is derived by:

$$\hat{p}_{\mathrm{MLE}}(c_j) = \frac{\mathrm{freq}(c_j)}{\mathrm{N}} \qquad (4)$$

where

$$\mathrm{N} = \sum_{j=1}^{116} \mathrm{freq}(c_j) \qquad (5)$$

and freq(c_j) is the frequency of the conceptual class c_j, which is defined as

$$\mathrm{freq}(c_j) = \sum_{w \in \mathrm{words}(c_j)} \frac{\mathrm{freq}(w)}{|\mathrm{classes}(w)|} \qquad (6)$$

words(c_j) is the set of words that belong to the conceptual class c_j, |classes(w)| is the number of conceptual classes of a word w, and freq(w) is the frequency of w.

The conditional probability of a particular conceptual class c_j given a predicate p_i is estimated from:

$$\hat{p}_{\mathrm{MLE}}(c_j|p_i)=\frac{\mathrm{freq}(c_j,p_i)}{\mathrm{N}} \tag{7}$$

where

$$\mathrm{N}=\sum_{j=1}^{116}\mathrm{freq}(c_j,p_i) \tag{8}$$

and

$$\mathrm{freq}(c_j,p_i)=\sum_{w\in\mathrm{words}(c_j)}\frac{\mathrm{freq}(w,p_i)}{|\mathrm{classes}(w)|} \tag{9}$$

words(c_j) is the set of words that belong to the conceptual class c_j, freq(w, p_i) is the co-occurrence frequency of the word w and the predicate p_i,[1] and |classes(w)| is the number of conceptual classes of w.

3 Disambiguation of VO- and VN-constructions

According to [2], ambiguities in V+N pairs are most difficult in transitive verbs. We therefore focus on disambiguating $\mathrm{V_{transitive}}$+N pairs; in particular, we focus on bi-syllabic transitive verbs. We extracted a total of 880 $\mathrm{V_{transitive}}$+N pairs from the Sinica corpus Version 1.0 [4]. 708 word pairs were used for training while the remaining 172 pairs were used for testing. The list of verbs covered are: 訓練 *xun3lian4* ‘train’, 表演 *biao2yan3* ‘perform’, 治療 *zhi4liao2* ‘cure’, 表達 *biao3da2* ‘express’, 學習 *xue2xi2* ‘learn’, 選擇 *xuan3ze2* ‘choose’, 生產 *sheng1can3* ‘produce’, 解決 *jie3jue2* ‘solve’, 教育

[1] In our experiment, the window size is set to 5. That is, a word w is regarded as co-occurring with a predicate p_i if it is not more than 5 words away from the predicate.

jiao4yu4 'educate', 發展 *fa1zhan3* 'develop', 處理 *chu2li3* 'handle', 參加 *can1jia1* 'participate', 管理 *guan2li3* 'manage', 建設 *jian4she4* 'build', 進行 *jin4xing2* 'go on', 使用 *shi3yong4* 'utilize', and 影響 *ying2xiang3* 'influence'. We manually separated the training set into VO-pairs and VN-pairs, from which we derived the selection preferences of each verb in a VO-relation and a VN-relation. The formulae used in the derivation have been covered in Section 2. These vectors of selection preferences (*$Pref_{VO}$* and *$Pref_{VN}$*,) are later used to determine whether a V+N pair is a VO-construction or a VN-construction. We will illustrate this step with an example.

Figure 2 and 3 show the selection profiles of the verb 影響 *ying2xiang3* 'influence' in a VO- and VN-relations respectively.

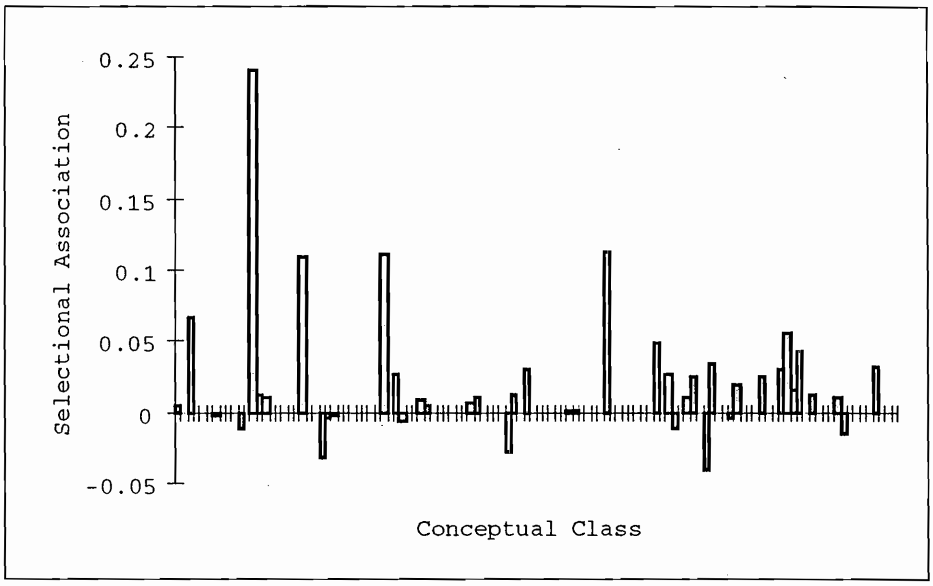


Figure 2. Selection profile of 影響 *ying2xiang3* 'influence' in a VO-relation

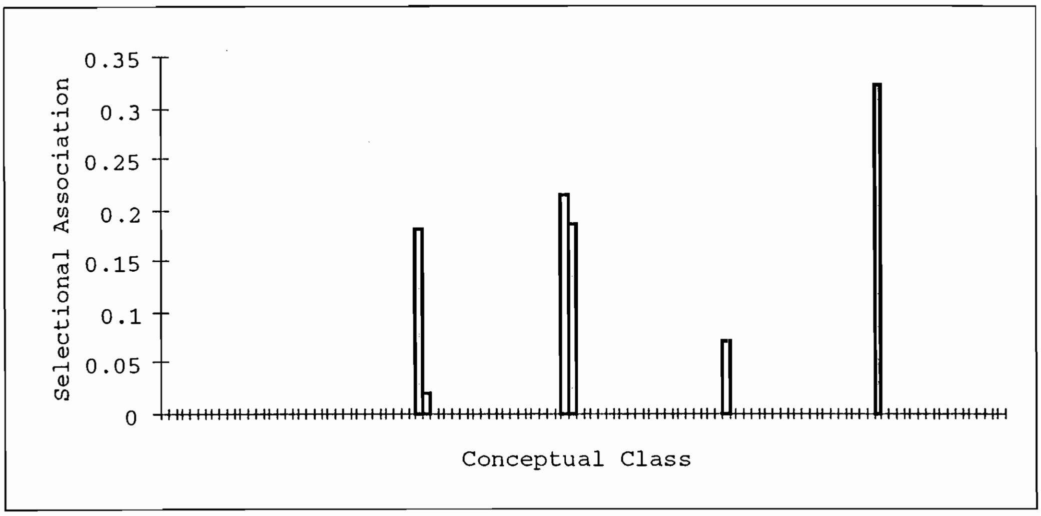


Figure 3. Selection profile of 影響 *ying2xiang3* 'influence' in a VN-relation

The selection profile of the verb 影響 *ying2xiang3* 'influence' in Figure 2 indicates that arguments of a wide range of conceptual classes could be the object of 影響 *ying2xiang3* 'influence'. On the other hand, the verb 影響 *ying2xiang3* 'influence' is modifiers of nouns of a much restricted set of conceptual classes (see Figure 3). During testing, say when 影響十成績 is presented, we first derive a binary vector *Q* that represents the conceptual classes of 成績 *cheng2ji1* 'results'. The vector is defined over the entire set of conceptual classes in a taxonomy (a total of 116 classes in our case). All conceptual classes that 成績 *cheng2ji1* 'results' belongs to will be assigned the value 1, with all the others 0. A **similarity measure** between *Q* and a preference vector *Pref* is defined as follows.

$$S = \frac{|Pref|}{M} Pref \cdot Q + T_V \cdot T_N \tag{10}$$

where $|Pref|$ is the total number of conceptual classes in *Pref* that have non-zero selectional association, and *M* is the total number of conceptual classes in a taxonomy. For the similarity measure of a verb *V* in a VO-construction (S_{VO}), T_V is a probabilistic measure of the tendency that the verb appears in VO-construction, and T_N is a probabilistic measure of the tendency that a noun appears in VO-construction. The former is called the **VO-tendency of a verb** (T_V^{VO}) while the latter is called the **VO-tendency of a noun** (T_N^{VO}). The VO-tendency of a verb *V* is estimated from the percentage of V+N$_i$ pairs in a corpus that are VO-construction. The VO-tendency of a noun *N* is estimated from the percentage of V$_i$+N pairs where *N* is the object of V_i. The **VN-tendency of a verb** (T_V^{VN}) and the **VN-tendency of a noun** (T_N^{VN}) in determining the similarity measure of a verb in a VN-construction (S_{VN}) can be derived from the following equations.

$$T_V^{VN} = 1 - T_V^{VO} \qquad (11)$$

$$T_N^{VN} = 1 - T_N^{VO} \qquad (12)$$

Our experimental data of the similarity scores of 成績 *cheng2ji1* ‘results’ with respect to the verb 影響 *ying2xiang3* ‘influence’ under VO-construction (S_{VO}) and VN-construction (S_{VN}) are 0.32 and 0.063 respectively. Since S_{VO} is greater than S_{VN}, we conclude that 影響十成績 is a VO-construction.

The ratio |Pref|/M in equation (10) is a weight used to implement the preference for a construction which has a selection profile that covers a wide range of conceptual classes. In the example of the verb 影響 *ying2xiang3* ‘influence’, the selection profile of the VO-construction is more spread out than that of the VN-construction. When a new pair of 影響＋N is encountered, assigning it as a VO-construction would have a higher chance of being correct. This heuristics is incorporated into the ratio. The additive term $T_v \cdot T_N$ incorporates the heuristics as observed by [2]: (i) When both the verb and noun in a V+N pair have a high VO-tendency, it is more likely that this is a VO-construction. Conversely, when both the verb and noun have a high VN-tendency, they are more likely to form a VN-construction. When the tendency of the noun and verb contradict each other, as well as when neither the noun nor the verb has a clear VN- or VO-tendency, selection preferences play a decisive role.

4 Experimental Results and Discussion

The experimental procedure is summarized as follows:

- extract from the Sinica corpus all sentences[2] which contain one of the 17 bi-syllabic, transitive verbs;
- extract semi-automatically all V+N pairs from the sentences found in step 1;
- manually split the V+N pairs into two groups: VO-pairs and VN-pairs;
- derive T_V^{VO}, T_N^{VO}, T_V^{VN} and T_N^{VN} in the manner as described in Section 3;
- use 80% of the VO-pairs and VN-pairs as training data to derive the selection preferences of each verb (see Section 2);
- use the remainding 20% to evaluate the performance of the proposed approach.

The similarity measure in equation (10) is used to determine whether a given V+N pair is a VO- or VN-construction. The decision is as follows:

[2] A sentence is defined as a sequence of characters delimited by punctuation marks.

```
if      $S_{VO} >= S_{VN}$
then    V+N is a VO-construction
else    V+N is a VN-construction
```

An average recognition rate of 88.4% is obtained in our experiment. A detailed break down is shown in Table 1.

Table 1. Recognition Rate of Each Verb

Verbs	**Recognition Rate(%)**
訓練 *xun3lian4* ‘train’	72.7
表演 *biao2yan3* ‘perform’	100
治療 *zhi4liao2* ‘cure’	100
表達 *biao3da2* ‘express’	80
學習 *xue2xi2* ‘learn’	71.4
選擇 *xuan3ze2* ‘choose’	100
生產 *sheng1can3* ‘produce’	66.7
解決 *jie3jue2* ‘solve’	66.7
教育 *jiao4yu4* ‘educate’	85.7
發展 *fa1zhan3* ‘develop’	91.7
處理 *chu2li3* ‘handle’	100
參加 *can1jia1* ‘participate’	94.1
管理 *guan2li3* ‘manage’	100
建設 *jian4she4* ‘build’	90
進行 *jin4xing2* ‘go on’	90
使用 *shi3yong4* ‘utilize’,	82.4
影響 *ying2xiang3* ‘influence’	95.5
Average	88.4

The VN-tendency of the 17 verbs are shown in increasing order in Table 2. The same table can also be interpreted as displaying the VO-tendency of these verbs in decreasing order.

Table 2. VN-tendency (in %) of the 17 Verbs

影響	8
選擇	18
處理	28
進行	29
使用	31
學習	31
表達	32
生產	35
發展	41
管理	43
參加	44
解決	46
建設	67
訓練	68
治療	68
表演	79
教育	84

Three factors that have impacts on the derivation of selection preferences are:

- **Word Boundary Accuracy** The accuracy of word boundaries in the Sinica corpus will directly influence the derivation of selection preferences. First, a sentence that contains a predicate p_i will be missed if the predicate is incorrectly segmented. Second, any error in the word boundaries of words that co-occur with the predicate will affect the estimate $\hat{p}_{MLE}(c_j|p_i)$ in equation (7). The Sinica corpus uses human labor to post-edit on the output of an automatic parts-of-speech tagger [6]. The post-editing work includes correcting errors in word boundaries and parts-of-speech [5]. In terms of word boundary accuracy, it is one of the best resources available currently.

- **V+N Pairs Extraction** Statistical techniques in general face the problem of insufficient data. Hence,the larger a test set is, the better are the statistical estimates. In this work, we used a semi-automatic approach to extract V+N pairs from the Sinica corpus. From all sentences that contain a particular verb, say 影響 *ying2xiang3* 'influence', we extract only those V+N pairs that are of these two patterns: V N+ and V N+ 的 N+.[3] In sentences (1) to (3), the followings: 影響十別人, 影響十世界, and 影響十形式 were extracted. We then manually went through all the extracted pairs to remove the erroneous ones and to decide whether they are VN- or VO-constructions.

(1) 即 卡耐基 的 那 本 《 如何
ji4 *ka3nai4ji1* *de* *nai4* *ben3* *ru2he2*
that is Carnegie DE[4] that CL[5] how
影響 別人 》。
ying2xiang3 *bie2ren2*
influence others
That is, the book "How to influence others" written by Carnegie.

[3] N+ refers to one or more nouns.

[4] DE refers to the structure word 的 de.

[5] CL stands for a classifier.

(2)	這些	都	不	足以	構成
	zhei4xie1	*dou1*	*bu4*	*zu2yi3*	*gou4cheng2*
	these	all	not	sufficient	constitute

衡量	一 位	影響	全	世界，
heng2liang4	*yi1 wei4*	*ying2xiang3*	*quan2*	*shi4jie4*
measure	one CL	influence	whole	world

These are not sufficient to measure a person who has influenced the whole world.

(3)	目的	在	了解	社會	因素
	mu4di4	*zai4*	*liao2jie3*	*she4hui4*	*ying1shu4*
	goal	at	understand	society	factor

如何	影響	語言	的	形式。
ru2he2	*ying2xiang3*	*yu3yan2*	*de*	*xing2shi4*
how	influence	language	DE	form

The goal is to understand how social factors influence language form.

Our simplistic approach inevitably leaves out many V+N pairs. Sentences (4) to (6) are some examples where this has happened. The object of 影響 *ying2xiang3* 'influence' in (4) is 日本 *ri4ben3* 'Japan', which appears in a passive sentence structure. Our approach missed this. In (5), the object has been wrongly identified as 部分 *bu4fen4* 'part' instead of 張力 *zhang1li4* 'tension'. In (6), 品質 *pin3zhi4* 'quality' instead of 生活 *sheng1huo2* 'living' should be the object of 影響 *ying2xiang3* 'influence'.

(4)	日本	受	儒家	影響。
	ri4ben3	*shou4*	*ru2jia1*	*ying2xiang3*
	Japan	receive	confucius	influence

Japan is influenced by confucius thinking.

(5)	減少	不	影響	張力	的
	ian3shao3	*bu4*	*ying2xiang3*	*zhang1li4*	*de*
	reduce	not	influence	tension	DE
	部分。				
	bu4fen4				
	part				

To reduce those parts which do not influence tension.

(6)	甚至	影響	生活	與	工作	的
	shen3zhi4	*ying2xiang3*	*sheng1huo2*	*yu3*	*gong1zuo4*	*de*
	even	influence	life	and	work	DE
	品質。					
	pin3zhi4					
	quality					

even influences the quality of life and work

- **Polysemy Issue** The Sinica corpus is not sense-disambiguated. Therefore, the selectional behavior of multiple senses of a verb is conflated. This is not necessarily a problem, as the resulting selection profile of the verb has distinct groupings. In determining the similarity measure using equation (10), only groupings that match the conceptual classes of a noun are considered.

The issue of polysemy also occurs in nouns. Our conceptual classes of nouns were based on the CKIP dictionary [3]. This dictionary has altogether 78,410 lexical entries, out of which 34,984 are nouns. The average number of senses per noun is 1.0115. Thus, most of the nouns in the CKIP dictionary have only one sense.

5 Comparison With Related Work

The approach described in [2] uses the following algorithm to decide whether a V+N pair is a VN- or VO-construction.

```
if      V is intransitive/pseudotransitive
then    V+N is a VN-construction
else    if V can be nominalized
        then    if V has a strong VN-tendency
                then    if   N is not an individuated noun[6]
                        then    V+N is a VN-construction
                        else    V+N is a VO-construction
                else    V+N is a VO-construction
        else    V+N is a VO-construction
```

Our work replaced the following steps of the algorithm by an information-theoretic approach as described in Sections 2 and 3.

[6] The followings are considered as individuated nouns: proper nouns, count nouns, location nouns, and pronouns.

```
if V has a strong VN-tendency
then  if    N is not an individuated noun
      then  V+N is a VN-construction
      else  V+N is a VO-construction
else  V+N is a VO-construction
```

It is not clear in [2] what threshold is used to decide whether a verb has a strong VN-tendency. The paper also did not explicitly state the performance of the algorithm. Thus, a quantitative comparison is not possible. Qualitatively, the approach in [2] uses the part-of-speech of nouns (i.e., whether a noun is an individuated noun) to decide whether a V+N pair is a VN- or VO-construction. Selection preferences in our approach in equation (10) is essentially a measure of the semantic compatibility between a verb and a noun. Our approach, in addition, has also incorparated syntactic factors. In [2], it is observed that individuated nouns usually have a strong VO-tendency while non-individuated nouns are more likely to have a strong VN-tendency. This insight on the syntactic behavior of nouns in V+N pairs is implicitly incorporated in the term T_N in equation (10). Individuated nouns will have a high VO-tendency value (T_N^{VO}) while non-individuated nouns will have a high VN-tendency value (T_N^{VN}). An advantage of our approach in comparison with the hard-and-fast rules in [2] is that exceptions to the rules can be handled better. For example, 影響十蹟象 will be identified as a VO-construction in [2] since the verb 影響 *ying2xiang3* ‘influence’ has an extremely weak VN-tendency (0.08 as shown in Table 2). The correct relation in this example should be a VN-construction, which is correctly identified in our approach because we consider not only the tendency of the verb, but also the tendency of the noun involved, as well as the selection preferences of the verb.

6 Conclusions

We have described in this paper a new approach to disambiguate VN- and VO-constructions using selection preferences. In addition to this semantic factor, our approach has also incorporated likelihood measures of the tendency of verbs and nouns functioning in VN- and VO-constructions. These measures are implicit syntactic factors. Our preliminary results based on 17 bi-syllabic, transitive verbs with a total of 880 V+N pairs is 88.4%. Our next goal is to evaluate this approach with a larger set of data.

Acknowledgments

We would like to express our thanks to the Academia Sinica, Taiwan, for giving us the permission to use the Sinica corpus and the CKIP dictionary.

References

[1] 莫若萍 (1992) 一個適用於剖析漢語的概念結構，中央研究院資訊科學研究所技術報告 92-04。

[2] 陳克健，洪偉美 (1995) 中文裡「動一名」述賓結構及「動一名」偏正結構的分析，第八屆計算語言學研討會論文集，1-13。

[3] 詞庫小組 (1993) 中文詞類分析（三版），中央研究院資訊科學研究所技術報告 93-05。

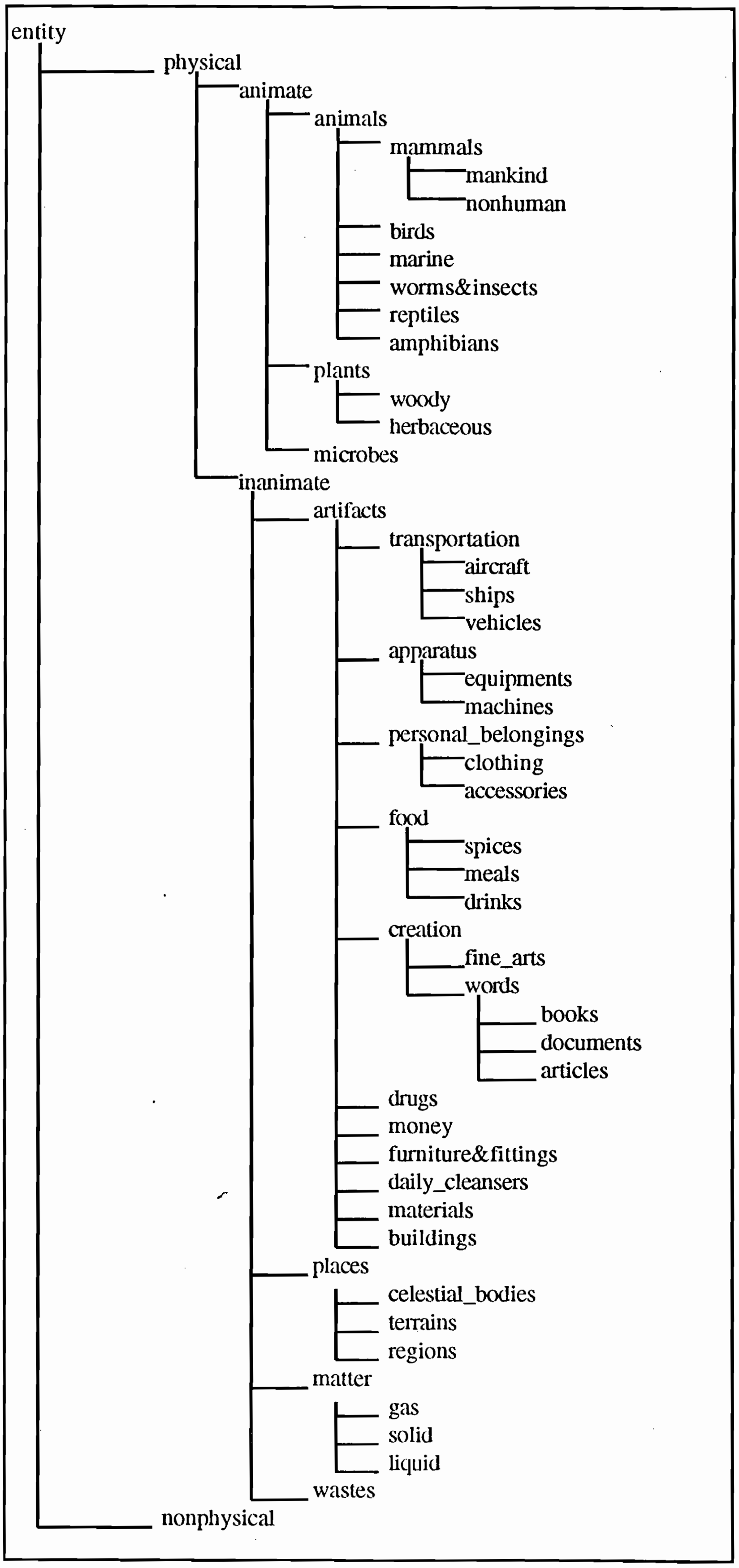
[4] 詞庫小組 (1995) 中央研究院平衡語料庫的內容與說明，中央研究院資訊科學研究所技術報告 95-02。

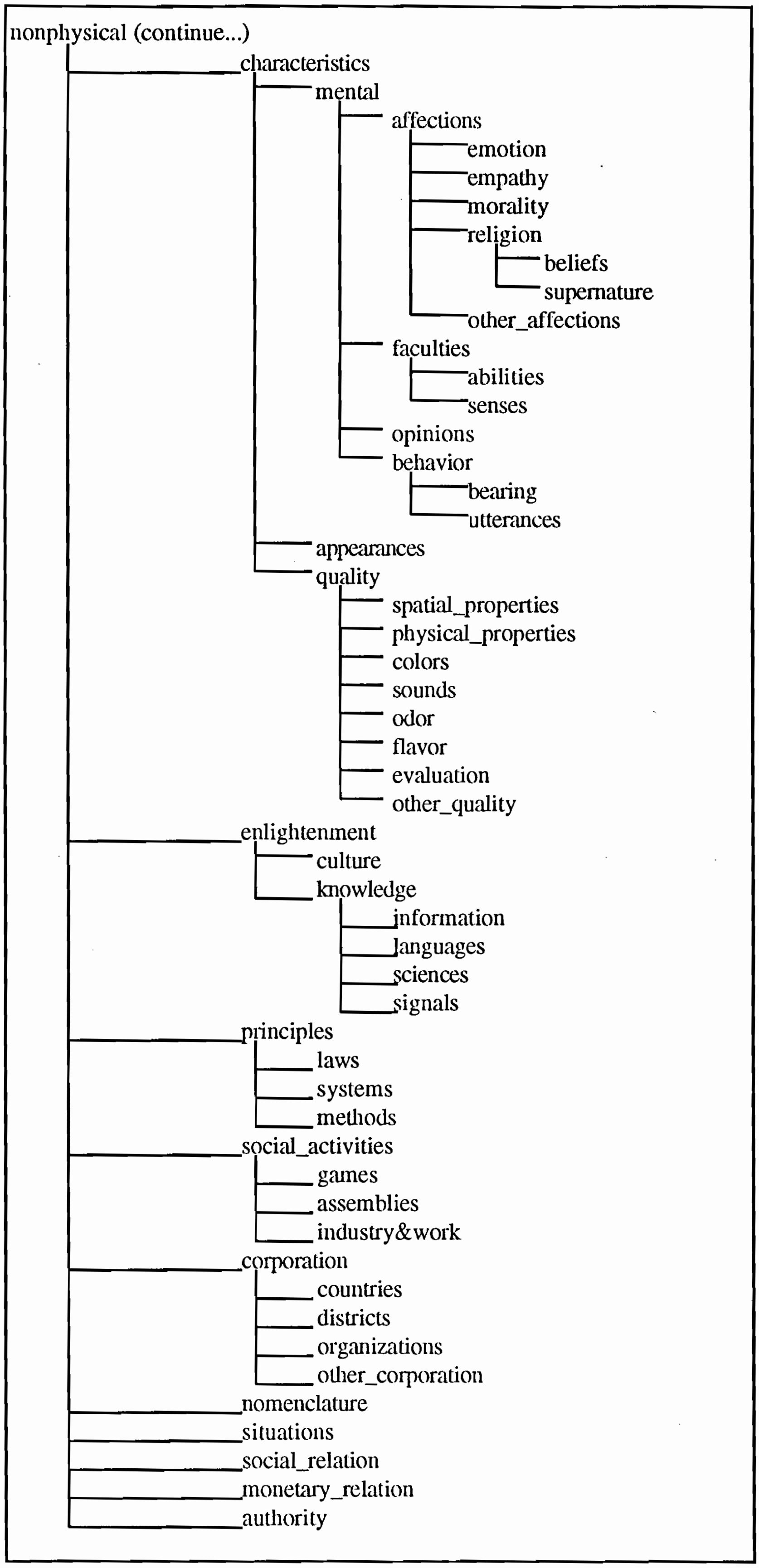
[5] Chang, Li-Ping, Chen Keh-Jiann (1995) “The CKIP part-of-speech tagging system for modern Chinese texts”. *Proceedings of ICCPOL 95.*

[6] Chen, Keh-Jiann, Liu Shing-Huan, Chang Li-Ping, Chin Yeh-Hao (1994) “A practical tagger for Chinese corpora”. *Proceedings of ROCLING VII,* 111-126.

[7] Resnik, P. S. (1993) “Selection and information: a class-based approach to lexical relationships”. Ph.D dissertation, University of Pennsylvania.

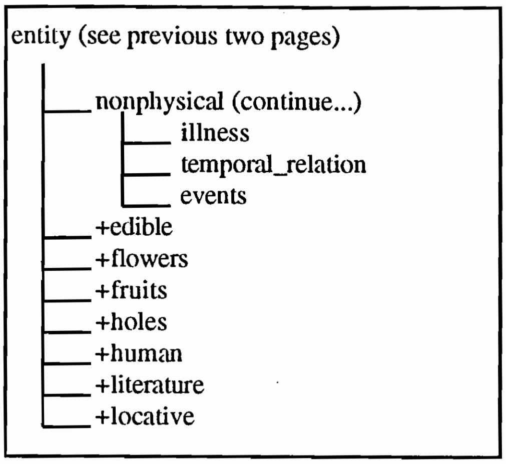
Appendix: A Taxonomy of Concept





nonphysical (continue...)

- characteristics
  - mental
    - affections
      - emotion
      - empathy
      - morality
      - religion
        - beliefs
        - supernature
      - other_affections
    - faculties
      - abilities
      - senses
    - opinions
    - behavior
      - bearing
      - utterances
  - appearances
  - quality
    - spatial_properties
    - physical_properties
    - colors
    - sounds
    - odor
    - flavor
    - evaluation
    - other_quality
- enlightenment
  - culture
  - knowledge
    - information
    - languages
    - sciences
    - signals
- principles
  - laws
  - systems
  - methods
- social_activities
  - games
  - assemblies
  - industry&work
- corporation
  - countries
  - districts
  - organizations
  - other_corporation
- nomenclature
- situations
- social_relation
- monetary_relation
- authority


entity (see previous two pages)
nonphysical (continue...)
illness
temporal_relation
events
+edible
+flowers
+fruits
+holes
+human
+literature
+locative