詞彙訊息的層次表達與管理

簡立峰　　國立台灣大學資訊系
陳克健　　中央研究院資訊所

摘要

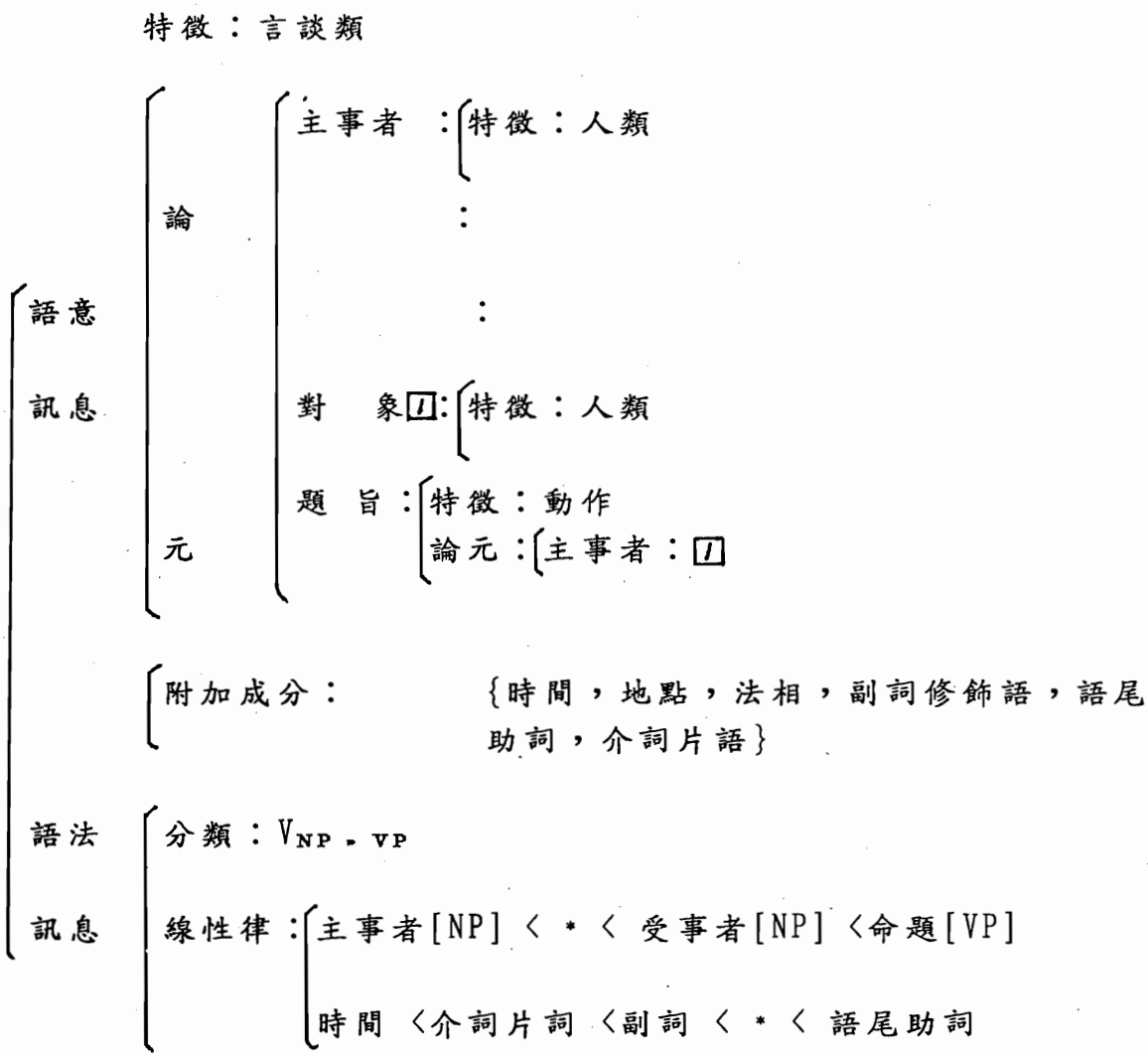
在這篇文章裡我們說明近來在自然語言處理所採行的語法訊息表達方式已有逐漸減少使用詞類訊息而改以加重詞彙個別特徵的趨勢。但是這類表達方式在實際運用上會導致(1)詞彙訊息建構不易，(2)訊息的正確性與一致性維持不易,(3)儲存空間耗費太大等問題出現。因此本文嘗試就這些問題加以探討，並且提出一套層次表達與管理方式藉以改善上述問題。利用層次表達具有以下的優點，1.保持資料的一致性，2.重覆訊息得以壓縮，3.訊息更動容易，4.有良好的訊息聚焦性，5.具彈性化的層次分類及訊息取得，6.有助於找出同類間細分類的準則，本文並討論了系統完成的方法。

1. 緒論

近代自然語言處理所採行的語法訊息表達方式，已有逐漸減少使用詞類細分類而改以加重詞彙個別特徵的趨勢。因爲一來個別特徵能更精確描述每個詞的特性，再者配合聯併語法[黃'88]的運用,這類訊息表達方式比傳統大量依賴詞組結構規律的表達方式會有更好聚焦性（執行效率較好）。但是在實際運用上,這類表達方式也會導致(1)詞彙訊息建構不易，(2)訊息的正確性與一致性維持不易，(3)儲存空間耗費太大等問題出現,如果這些問題無法加以克服,將使得依賴這種訊息表達方式的語法模式在實際運用上滯礙難行，因此本文嘗試提出一套層次化的詞彙訊息表達以及管理方法，希望藉此可以改善上述問題。

傳統上詞彙的語法特性都是依靠詞類分類以及詞組結構律來表達，例如，'打'在分類上屬於及物動詞Vt，而及物動詞的語法特性就是當形成動詞片語時需要後一個名詞片語,即結構律VP→Vt NP所表示，因此'打'的語法特性即可以Vt的特性來代表。雖然這種表達方式頗爲精簡，但也常常面臨一個難題，即同類的詞彙常未能具有完全相同的語法特性，因此通常得將這些詞彙再加以細分。如同生物類上有界、門、綱、目、科、屬、種等層次，在詞類分類中也會有層次的關係,例如一般動詞可分及物,不及物。而及物動詞可再分爲單及物或雙及物，甚至進而再細分可還因賓語的不同而區分爲接受名詞、動詞片語或句子的及物動詞。由於性質完全相同的兩個詞很難存在，就如同你、我被歸類於人科人屬人種，但你、我之間還是會有一些細微區別。對同類的兩個詞而言，一樣會有語法特徵上的差異。

爲了避免傳統這種以分類加詞組律的描述方式會有如 (1)只能表示分類的共同訊息，個別差異無法修正，(2)一致性與共存限制等問題必須以相似規律重覆表示， (3)無法兼顧語意的訊息，(4)表示法不夠精簡，對於詞序比較自由的語言，表達時更爲繁瑣,(5)因爲分類過細而造成較差的聚焦性等缺陷[陳'89]。近代語法理論上，多只選擇最重要的語法特徵加以分類，而對於一些次要或不適合分類的特徵（例如單、複數）多改爲以附帶特徵的方式來表達而不再以共同訊息視之。因而在強化詞組語法架構下(Augmented Phrase Structure Rules)會有以詞類+特徵的方式來表達語法訊息[Gazdar'85]。在這種架構下，詞類共同特性仍是表現在相關的詞組律上，至於同類詞彙間的差異則以個別特徵視之。此外，還有一些語法表達架構，例如HPSG[Pollard'87]，Categorial Grammar[Uszkoreit'86]甚至有逐漸簡化詞類特徵,轉而注重個別特徵的傾向,其中HPSG將詞組律減至數條而 Categorial Grammar則從根本上以類別取代了詞組律的使用。近來，[陳'89]更提出一個訊息爲本的格位語法表達模式(ICG)，其假設每個詞彙所有的重要語法、語意訊息都完全表達在特徵結構上，藉此詞組律將可完全捨去，如圖一所示動詞"勸"的例子。



圖一　詞彙'勸'的特徵結構

述訊息表達方式的優點是利用個別特徵可以更精確描述每個詞的特性，而詞組律的減少也使得聚焦性變得更好（畢竟在傳統一個有數百條詞組律的系統裡，與某一個詞類有關的規律一定十分有限，其聚焦性自然較差，這也是爲什麼會有 LR Parsing Table— 一種在剖析時協助聚焦的方法出現[Tomita'85]）。圖二進一步表列了上述有關敘述。

相關語法模式	儲存在詞彙各別詞項之訊息	以詞組律表達詞類分類訊息的情形
PSG	詞類別	大量的詞組律
GPSG	詞類別+個別特徵	強化詞組律
HPSG	以複雜特徵結構表示	
	所屬類別訊息+個別特徵	少數詞組律
Categorial	"	無
Grammar		
ICG	"	無

圖二　相關語法模式的詞彙訊息表達方式

然而儘管上述這些注重詞彙個別特徵的語法表達架構有較多優點，但是在實際應用上這種表達方式也衍生不少問題。如因儲存詞彙個別特徵所大量增加的空間需求絕不是一般小型電腦所能負荷。還有個別詞彙所應儲存訊息多半需要異動。然而異動的影響常常導致相同性質的詞彙會有不一致的訊息，例如，我們希望在所有的及物動詞的賓語中增加一語意限制，就必須修改所有及物動詞的詞項。可是萬一有某個詞彙沒有修改到，則除非在日後藉由聯併某些句子而得知，否則這種不一致是很難被發現。這樣自然使得詞彙訊息的構建變得非常困難。因此在這種處理環境裡我們是不能直接把詞彙訊息完全儲存在詞典的各別詞項底下而不加管理，我們得尋求其它更理想的表達與管理方法。因此本文嘗試提出一套層次化的詞彙訊息表達及管理方法以改善上述問題。目前這套方法已被用在建構根據ICG所編定的中文詞檔維護系統。在以下幾節裡，第2節介紹以層次結構來表達詞類訊息。第3節則提出以編輯記錄的方式表示詞彙與其所屬詞類差異。第4 節則是系統製作上的其它考。最後，第5節是結論。

2. 層次化詞彙訊息表達

根據上節所述，在注重詞彙個別特徵的自然語言處理環境下，直接將詞彙所有的訊息完全儲存在詞典個別的詞項底下基本上是不可行的，必須有更

好表達方式來替代。由於即使詞彙所有的語法、語意訊息都以複雜特徵結構表示，但在同類的詞彙間還是會有相同的訊息。因此詞類訊息的區分在這種環境下依然是很重要的。我們可以以如下的關係表示：

完整的詞彙訊息=同類詞彙間的共同訊息(即所屬詞類訊息)+
同類詞彙間之差異訊息(即與所屬詞類訊息的相對差異)。

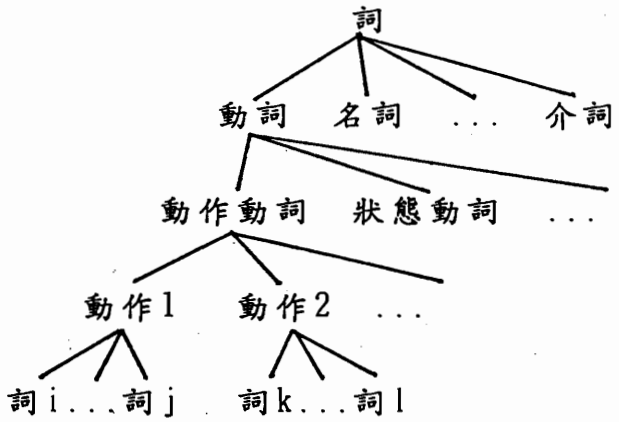
以這種方式表達詞彙訊息在系統建構以及儲存上將會方便且有效率多。因爲對於每個詞彙而言，建構訊息所需的動作只是擇定所屬類別以及列出與此類別所屬特徵結構的差異即可。例如中文詞'勸'在ICG裡屬$V_{NP,VP}$類，其特徵已如圖一所示，由於$V_{NP\cdot VP}$的特徵結構（如圖三所示）與'勸'頗爲近似，因此在勸的詞項下只要表示出與$V_{NP,VP}$的差異即可，而不必儲存整個訊息(表示差異的方法將在下節裡說明)。

- 語意訊息
  - 論元：
    - 主事者：
      - 意義：?
      - 特徵：?
    - 對象：
      - 意義：?
      - 特徵：?
    - 題旨：[ ]
  - 附加成分：{時間，地點，法相，副詞修飾語，語尾助詞，介詞片詞}
- 語法訊息
  - 分類：$V_{NP\cdot VP}$
  - 線性律：
    - 主事者[NP] < * < 受事者[NP] < 命題[VP]
    - 時間 < 介詞片詞 < 副詞 < * < 語尾助詞

圖三　$V_{NP\cdot VP}$的特徵結構

此外由於詞類分類，如前所述不論何種語言都會有大分類、細分類、再細分類等。如在ICG 內、其大分類包括中文三大詞類介詞、名詞、動詞,而動

詞二類、…，等。因此這些類別很自然地形成如圖四所示的層次結構關係。



圖四 ICG的詞類分類舉例

事實上我們可以以這種層次結構來表示詞類訊息，而個別詞彙的訊息即可視爲是儲放在這個層次結構的最底層。由於在層次結構裡，上層節點與直屬下層節點有一種 is- a關係，這種關係具有特性承襲的性質(Inheritance Property)。所謂特性承襲事實上就是把高層次的訊息聯併給低層次的子孫。這種結構所具有的優點包括訊息分享（上層訊息對下層節點而言可以直接繼承而不必重建)和一致性維持(任一節點的訊息異動都可傳承給下層子孫而不會有不一致的現象）。我們可以利用這種層次結構來表示詞類訊息，讓每一個詞類在結構上都有一節點，而這詞類的訊息即是由聯併其所有祖先節點的訊息以及本身節點的訊息而成。

因爲詞類的訊息較易整理，利用這種層次結構對訊息分享（減少儲存空間，表示彼此的相同特性)以及一致性的維持(可以修改上層節點的訊息）都有助益。此外，因爲下層的特徵值有較高的優先性 (priority),如果有例外的情形還可加以修正 ,例如企鵝是鳥類，企鵝承襲了鳥類的所有特徵，包括會飛的特徵。但是企鵝實際上不會飛，因此只要在企鵝的特徵裡註明會飛，則由於詞彙特徵是以下層爲主，就可以簡單的把例外情形解決。

綜合上述層次化訊息的表達方式具有下列的優點：

綜合上述層次化訊息的表達方式具有下列的優點：

(1) 資料的一致性(data consistancy)

由於訊息具有承襲的特性，同類詞彙的語法特性若欲修改，只要修改上層的類別訊息，不必一一更改同類棟的每一個詞彙資料，不致產生掛一漏萬的情形。資料維護容易，也可以保持正確性及一致性。

(2) 資料的壓縮(data redution)

由於同類的詞彙訊息,只在高層的類別訊息上表示一次,減少了許多不必要的重複訊息，達到了分類的主要目的。

(3) 訊息更動容易

修改或增加訊息可以適切的選擇正確的層次加以更動。其作用是全面性的更動了層次以下的所有詞彙訊息。

(4) 訊息的聚焦性

從層次結構上很容易獲得每個詞彙的完整訊息，摒除了其它無關的訊息。對剖析器而言，只須處理相關詞彙的語法語意訊息。可以減少許多不必要的檢測。

(5) 彈性化的分類及訊息取得

層次結構代表的是類別及細分類關係，對不同的應用而言，可能要求的訊息詳細程度有所不同。可以彈性的選取至不同的層次，此外分類也較爲容易。

(6) 有助於找出同類間細分類的準則

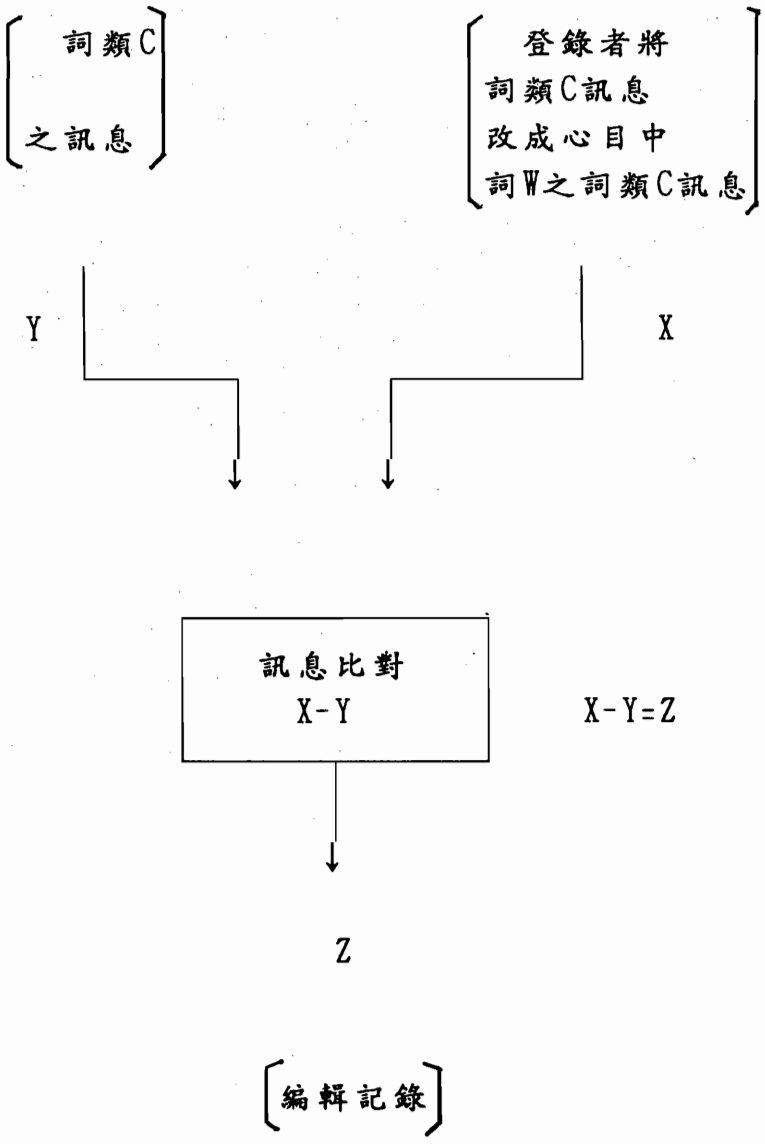
要找出細分類的標準通常並不容易，必須非常小心的觀察同類元素間的相同、相異點、找出足以作爲細分類的特徵。而這些細分類的特徵卻很容易的顯現在同類詞彙的個別差異訊息當中。

3. 編輯記錄與詞彙個別差異

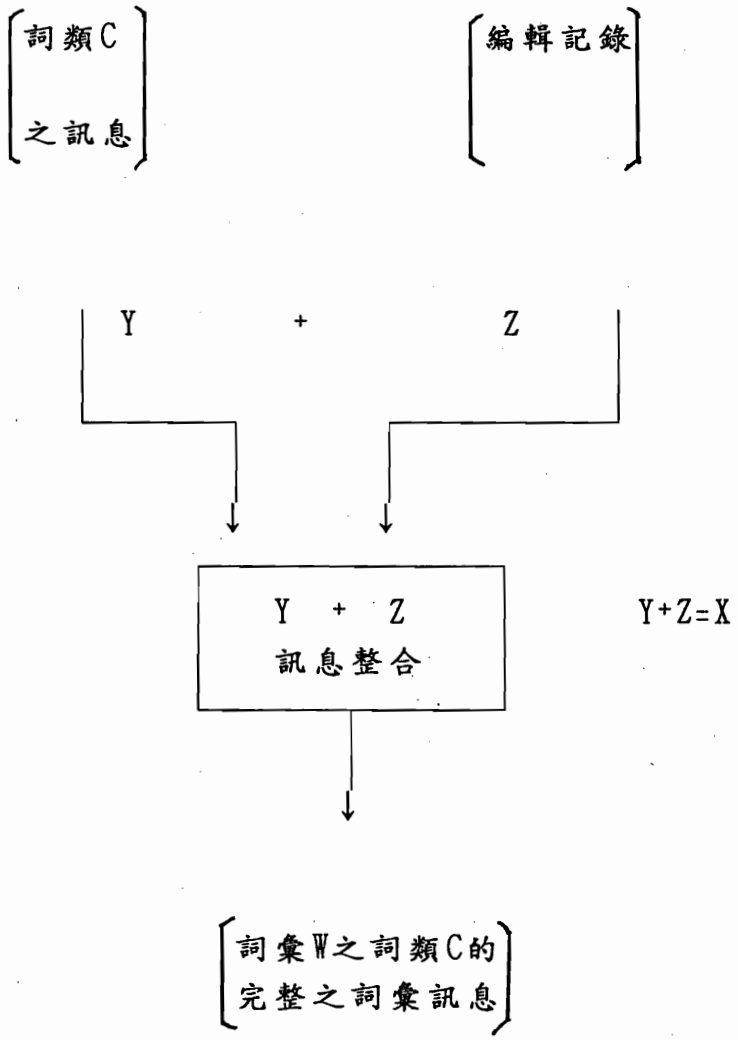
在上一節裡我們介紹以層次結構來表示詞類訊息，但是表達一完整的詞彙訊息還得包括相對於所屬詞類的差異訊息。然而這個差異訊息如何建立呢?事實上對訊息登錄人員而言要他們在詞項下描述與所屬詞類訊息的差異並不容易，例如填上類似" '勸'的主事者必須具有人類的特徵值，但類別 $V_{NP,VP}$

內則無此限制"這樣的敘述並不容易。但是如果將 V_{NP} - VP拷貝一份讓他們利用編輯器改成心目中'勸'的特徵結構，他們似乎很容易地就會在 V_{NP} - VP的主事者特徵後面加上了人類這個特徵值。於是我們進步提出編輯記錄(Editing Record)的表示方法來使得系統得以自動地表達出這種差異。這個處理方式可以如圖五所示：

(i) 詞彙訊息構建



(ii)詞彙　　訊息整合



圖五　編輯記錄和詞彙訊息構建與整合

圖五(i)所示是資料登錄者初次建構詞W之詞類C的訊息所採行的流程。圖五(ii)則是日後異動或審視時的流程。所謂編輯記錄Z即是登錄者心中認爲的完整詞彙訊息X和其所屬詞類C之訊息Y的差異 (Z=X-Y)。不論訊息構建或整合,對於登錄人員而言會以爲儲存在詞檔裡的訊息即是X，然而事實上卻是Z。這種由系統自動產生差異的方式會使得登錄人員感到非常方便而且系統效率也很好。然而編輯記錄是什麼形式的資料且系統又如何產生?事實上在ICG詞檔

維護系統裡不論詞類或詞彙訊息都被視爲一正規語言，有文法來描述以及剖析器來檢錯並將訊息轉換成串列的形式，因此圖五(i)中的 X和Y即是經由剖析產生的串列，至於X-Y即是X,Y二串列的比對，而最後的Z 就是有關X，Y二串列差異的描述，即所謂的編輯記錄。這類編輯記錄的每一筆格式如下：

路徑名/動作/字串
如果前述C代表如下特徵：

```
[sem: [arg: [agent: ]
      [adjunct:
```

而登錄者將其修改爲如下W的特徵：

```
[sem: [arg: [agent:[sem:human]
      [adjunct:
```

則Z就是如下之記錄：sem.arg.agent/Insert/sem human

也即在C 中Agent的特徵下增加一 "人類"語意特徵限制。而最後儲存放在詞典W詞項下的訊息將只有類別名稱C和編輯記錄Z(以代碼方式儲放)。

其實這種編輯記錄的最大好處還是訊息一致性的維持。例如若我們希望在所有 $V_{NP\ .\ VP}$類的詞彙增加一個特徵。我們只需在 $V_{NP\ .\ VP}$ 類別訊息加上此一特徵即可。因爲當初建構詞彙時這個特徵並不存在 $V_{NP\ .\ VP}$ 中，所有屬於這個詞類的詞彙其編輯記錄內都不會有任何有關此特徵的動作(也即沒什麼不同)，因此這個訊息便可輕易地就加入所有這類詞彙中。若同樣地日後我們要除去此特徵，只要將其從 $V_{NP,\ VP}$中除去即一樣地等於所有有關的詞彙內也同時刪除。

4. 製作上的其它考量

綜合上述介紹，我們以層次結構來表達詞類訊息，因而在系統製作上有了詞類訊息管理子系統的需求。加上將個別差異訊息 (其實是編輯記錄)儲放在詞彙個別詞項,因而有了詞彙訊息管理子系統。然而成爲一個完整的詞檔維護系統,在ICG的詞檔維護系統裡我們還做了如下的考量：

a.運用詞樹與直接

檔案取存技術快速存取訊息由於詞彙眾多且詞彙資料變動長，因此訊息存取必須有相當之技巧，我們利用變動長直接檔存取技術(direct file access with variable record size)和詞樹 (word tree)配合hash searching之應用使得訊息檢索之存取相當便捷。

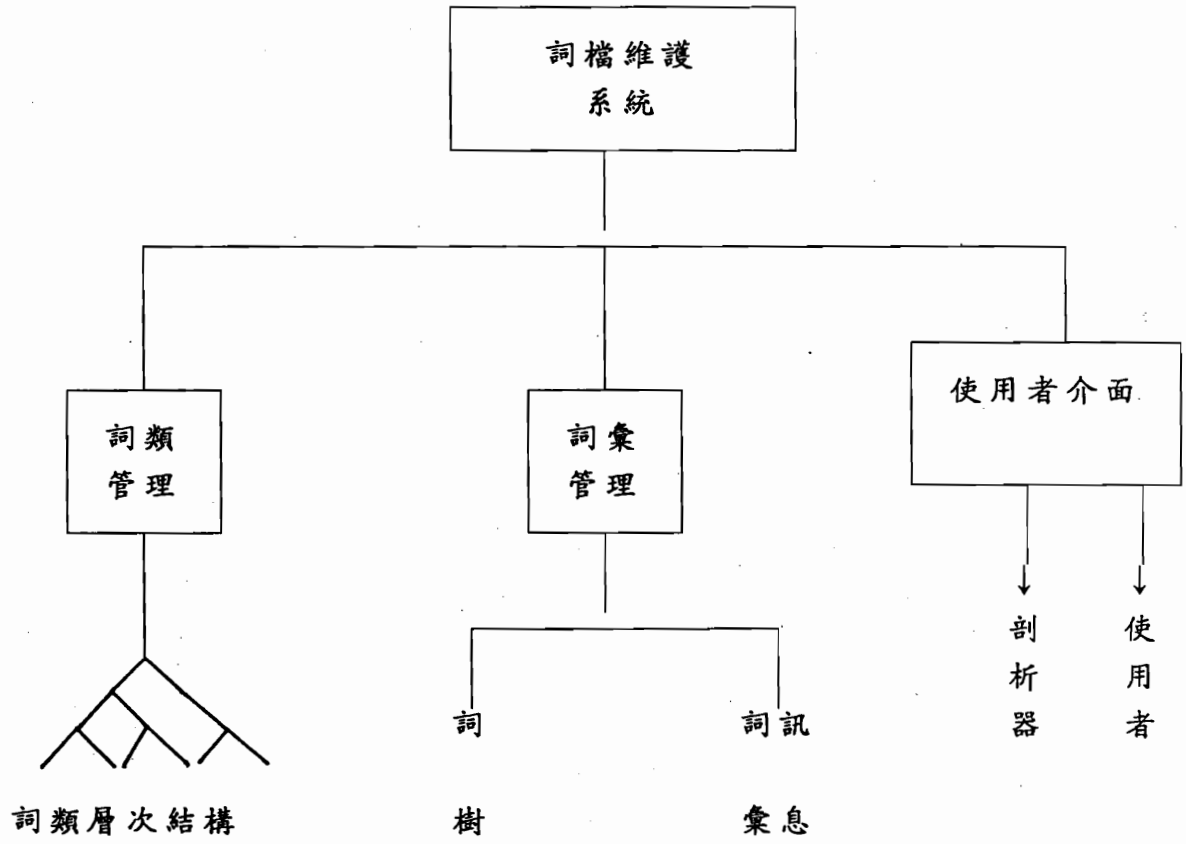
b.視訊息爲一種正規化語言

詞彙訊息必須能夠被剖析器解釋，而且詞彙訊息的正確性也必須加以維護。所以詞彙訊息絕對是一種有規律的知識，而非雜亂無章的資料。因此我們將訊息視爲如程式語言一樣的正規化語言，我們訂定一文法來描述，進而可以利用此文法來剖析資料之正確性。在詞類與詞彙兩個子系統裡我們都利用文法來協助登錄並檢錯。

c.目標導向式系統設計

在本系統內每個詞彙、每個詞類的每一個單項訊息都被視爲一目標(object)，每個目標有其個別的操作動作，如存取、編輯。因此系統設計上非常方便，擴充性也極強。

目前我們完成的管理系統其結構如下圖：



圖六　ICG詞檔維護系統架構

這個系統現在已開發完成並且開始建構詞彙訊息。

5. 結論

在本文裡我們首先提出近來在自然語言處理所採行的語法訊息表達方式已有逐漸減少使用詞類分類訊息而改以加重詞彙個別特徵的趨勢，然而這類表達方式在實際運用上會導致(1)詞彙訊息建構不易，(2)訊息的正確性與一致性維持不易，(3)儲存空間耗費太大等問題出現。因此我們提出利用層次結構來

表達詞類訊息，以及編輯記錄來表示詞彙個別差異。因爲利用層次結構管理詞類訊息這種複雜特徵結構，對儲存空間節省與詞彙分類都有很好的效果。加上以編輯記錄的方式來表示詞彙與其所屬詞類訊息差異的情形，可使得訊息構建自然且容易得多;另外正確性與一致性的維持也可以達成。有關這些處理方式我們已成功的運用在構成ICG詞檔的系統製作,此外我們還加上一些考量使整個系統非常有效率，我們相信這套詞彙訊息表達與管理方法對其它性質相近的聯併語法理論而言，都有進一步參考的價值。

6.參考資料：

1.[黃'88] 黃居仁，聯併：語法理論與剖析，ROCLING'88

2.[陳'89] 陳克健、黃居仁,訊息爲本的格位語法— 一個適用於表達中文的語法模式，ROCLING'89

3.[Gazdar'85]
Gazdar. G.,E. Klein, G.K. Pullum,and I. A. Sag 1985. Generalized Phrase Structure Grammar. Cambridge: Blackwell, and Cambridge, Mass.: Harvard University Press.

4.[Pollard'87]
Pollard,C.and Ivan A. Sag 1987. Information-based Syntax and Semantics Vol.1 Fundamentals, CSLI Lecture Notes Series No. 13. Stanford: CSLI.

5.[Tomita'85]
Tomita,M. 1985, Efficient Parsing for Natural Language, Kluwer Academic Publishers.

6.[Uszkoreit'86]
Uszkoreit, H. 1986. Categorial Unification Grammars. In Proceedings of Coling 1986. Bonn University of Bonn. Also appeared as report No. CSLI-86-66, Stanford: CSLI.

*本論文的研究得到國科會中文語句剖析的語法模式研究計畫(NSC78-0408-E001-01)及中央研究院與工研院電子所合作「中文剖析系統合作研究發展計畫」(X2-79007)之部分經費補助，特此申謝。此一層次維護系統的程式設計員有張孟元、蔡正茂、莊清華及溫佩樺，謝謝他們對本篇論文的幕後貢獻。論文寫作的部分例子取材於中文詞庫小組的研究成果，並感謝詞彙小組對此一層次維護系統所做的實際測試。