計算機處理自然語言之研究與語法理論及其形式組之應用

語言學研究所副教授　何萬順

ABSTRACT

Research in natural language processing (NLP) is concerned with the implementation of linguistic theories and formalisms in computational systems processing human languages. This paper first examines the significant place a proper linguistic theory and its corresponding formalism hold in NLP research. We then evaluate how the mainstream transformational theory of Government and Binding (GB) compares with the lexically-oriented non-transformational Lexical-Functional Grammar (LFG) in its computational applications in NLP. Within relevant contexts we will also touch upon actual NLP projects that deal with the

Chinese language. Most importantly we will attempt to demonstrate the suitability of the formal theory of LFG, especially as expressed via the vLFG formalism developed in Her (1990), for NLP applications such as sentence analysis, generation, and machine translation.

一、背景說明

自然語言處理（natural language processing）簡稱為NLP，溶合了語言學及計算機學的專門知識。以計算機來處理自然語言包括了對語音、詞匯、句法、語意、語用、及言談篇章各層面的分析。本文討論的重點在於語法（grammar）因此將不涉及語音、語用及言談篇章。從語法來看，NLP的目的不外是用計算機來剖析、生成、及翻譯語言，而在設計一個NLP系統時的第一步，就是選擇一個適當的語法理論（theory）和一個相對的形式組（formalism）。這個理論要足以對語言的各種現象提出正確的解釋，而其形式組則必須提供精確的形式表現，適於計算機的運用。

本文首先說明語法理論、形式組、及NLP的不同功能及目的，探討三者在研究發展時相互依存的密切關係。然後我們對主流的變換理論：管轄約束語法（Government and Binding）簡稱GB，和一個以詞匯為核心的非變換理論：詞匯功能語法（Lexical-Functional Grammar）簡稱LFG，就二者在NLP的應用上進行評估。本文的重點在於顯示LFG語法理論及vLFG-F形式組適用於計算機處理語言的運用，無論是在剖析（parsing），生成（generation）、或是機器翻譯

（machine translation）的功能上，LFG 的運用都要比 GB 更有效率。

二、語法理論、形式組、及自然語言處理

在衍生語言學（generative linguistics) 的傳統裡，杭士基（Chomsky 1957）所堅持的，語法理論的目的在於界定所有的可能的自然語言並且排除不可能的人類語言，因此對語言的本質提出假設，並且對各個獨立語言中的語法結構提出分析（analysis）。形式組（formalism）的目的是在一個語法理論的許可範圍內，對其語法分析提供一個確切的形式的表現。可是正如希伯（Shieber 1987）所充分顯示的，一個語法極少是僅能由某一個形式組來表達；在同一篇文章，希伯對於語法理論、形式組、及語言分析之間的關係有以下的圖解（Shieber 1987:5）。

THEORY	Example： LFG
↓ restricts（限制）	
FORMALISM	LFG formalism
↓ expresses（表達）	
ANALYSIS	English
↓ predicts（預測）	
ANALYSES	English sentences

圖一、語法理論、形式組、及分析之關係

NLP 的目的既然是以計算機來處理語言，語法理論自然在 NLP 研究發展上有中心地位；所有的計算機系統必須為形式系統，因此語法理論的形式組也是 NLP 研究上不可或缺的一部份。然而，一個語法理論所允許的形式組中很可能在 NLP 的應用上有優劣之分；正如不同的語法理論在 NLP 的應用上有優劣之分一樣。本文的重點正是在於顯示在 NLP 的剖析、生成、和機器翻譯數種功能裡，詞匯功能語法（ LFG ）的應用要優於變換語法 GB；再者，我們要顯示 LFG 理論下的兩個形式組，一是傳統的 LFG 形式組，我們以 LFG-F 來代表，另一個是作者與 Joseph Pentheroudakis 及 Dan Higinbotham 所發展出的 vLFG 形式組，以 vLFG-F 來代表，後者在 NLP 的應用要優於前者。

1.管轄約束理論與詞匯功能語法

杭士基現今的管轄約束理論（ GB ）仍然被視為是主流的語法理論，承襲了他早期的變換語法理論（ Transformational Grammar ）。然而不以 GB 為理論基礎的語法研究者其實才是多數；更重要的是新的語法理論興起並且有些也已經廣為語法學家所接受，參見 Sells 1985, Horrocks 1987, 及 Shieber 1987。此外，因為計算機學的發展，語法理論在計算機的應用上得到實証以及實用；在這一方面 GB 已失去了主導的地位。早期的變換語法在計算機應用上曾遭遇嚴重的困難，而 GB 的語法研究者對於形式表現的精確也不再加以重視，這些都導致了 GB 在 NLP 應用上的式微。

詞匯功能語法（ LFG ），與 GB 相比之下，積極地應用於 NLP 的研究中，因此在理論與實用之間達到相輔相成的作

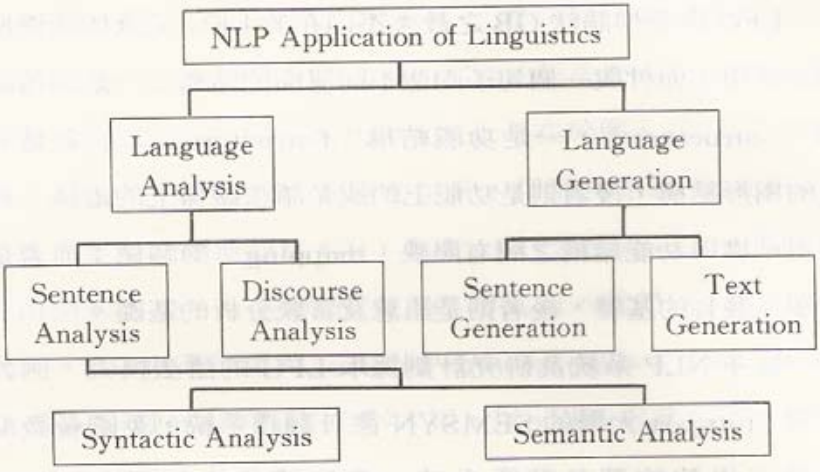
用。其他的聯併語法（Unification Grammar）在形式上與LFG相似，這些理論及形式組的興起以及廣泛使用也助長了LFG的普遍性。

LFG與變換語法GB之最大不同在於LFG完全排除變換律的使用，而付與一個句子兩個不同層面的結構：一是詞組結構（c-structure），一是功能結構（f-structure）；前者是一般的樹形結構，後者則是功能上的或是語法關係上的組織。在詞組結構與功能結構之間有照映（mapping）的關係；前者是音韻及發音的基礎，後者則是語意及言談分析的基礎。

許多NLP系統及研究計劃採取LFG的語法模式，例如德國Stuttgart大學的SEMSYN德日翻譯系統，英國曼徹斯特理工學院的英日翻譯系統，還有著名的美國Carnegie Melon大學的以知識為基礎的KBMT翻譯系統（Carbonell and Tomita 1987）。甚至歐洲共同市場所資助開發的EUROTRA計算機翻譯研究計劃也有專家建議改用LFG為語言分析的基本模式（Gebruers 1989）。美國ECS公司的研究計劃以LFG為核心，也已成功地研發了五個不同語言組的翻譯機（Her 1989）。LFG的聯併形式也表示它與其他以聯併為主的語法之間的互通性，如功能聯併語法（Functional Unification Grammar, FUG）（Kay 1986）及PATR II（Shieber 1986, 1987）。而共通性也是在進行NLP研發時的一個實際的重要考量。LFG因此無論是在理論上及應用上均已獲得肯定。

2.自然語言處理的分類

在 NLP 的研究中，依語法的角度來看可以做出以下的分類：



圖二、語法在 NLP 研究中的應用

句子分析一般稱作剖析（parsing）。句子生成（generation）有時也稱作組合（synthesis）。

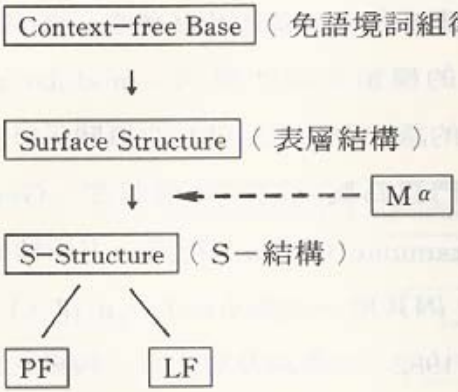
GB 由於變換律的緣故，句子的生成和剖析的運作在過程上差異很大，剖析的複雜度要遠超過生成。在剖析時要恢復變換律前的深層結構（d-structure）是一大難題（King 1983）。由於 X-bar 的部份只生成深層結構，表層結構（s-structure）並不在 X-bar 的限制之下，因此表層結構因為有了諸如移動（movement）的變換，結構大有改變。因此 GB 的生成語法不足以直接使用於剖析。

此外，語法理論和其形式組在理想上應該要能反應人類處理語言的心理過程。例如，Crain 和 Fodor（1986）認為實驗顯示人類剖析語句時，詞組結構和空填關係（filler-gap

dependencies）的限制兩種訊息是同時存在的；根據這個發現他們發展類似 GB 的模組式語法模式（modular grammar model），因為這樣的語法中詞組結構和空填關係限制分屬兩個不同的模組；他們認為概化詞組結構語法（Generalized Phrase Structure Grammar, GPSG）在這一方面精確地反映了人類的句子剖析，因其單一一致系統的詞組律。LFG 如何呢？根據 Bresnan（1982）的理論及形式組，詞組結構完成之後才產生功能結構，而空填關係的限制是在功能結構之內；因此，LFG 似乎也無法正確地反應這個心理過程。然而，在 vLFG-F 的形式組之下，詞組結構產生的同時也必然要有其功能結構，因此在這個形式組之下，詞組訊息和空填關係限制的訊息是同時並存的。因此能反映此一心理過程。

三、自然語言剖析

在計算機剖析自然語言的應用上，LFG 和類似的聯併語法在近年來是最廣為接受的。GB 和類似的變換語法則應用較少，一大原因就是因為變換律在剖析上造成的困難。以 GB 為基礎的剖析器（parser）必須先設定一組表層詞組律以決定句子的表層結構，然後再反轉（reverse）變換律（如 Move-α 及其他刪減律）的效果，最後才得到深層結構。比起 LFG 的剖析過程，GB 剖析要繁複許多。因為如此，Correa (1987) 所設計的 GB 剖析器採用了一個非變換的 GB 模式；此一模式 Chomsky（1981:89-92）曾略為提到。我們將這個非變換的 GB 模式以圖式列出，其中 LF 是 logical form（邏輯形式），PF 是 phonological form（音韻形式）。



圖三、非變換的 GB 語法模式（Correa 1987）

由上圖可見在 Correa 的剖析器裡 Move-α 的變換律已由 M_α 的解釋系統（interpretive system）所取代，而所謂的深層結構則完全不存在。Chomsky 本人雖然了解這個非變換模式的可能性，但他並不認為這和變換模式之間的不同有任何的重要性（Chomsky 1982:33）。從純理論的角度以及衍生能力（generative power）來看，Chomsky 是正確的，但是在計算機實際應用上，Correa 所採用的形式模式卻方便了許多，因為在表層結構與 s-結構之間，詞組結構完全相同，因此在剖析上只需要一組的詞組律，兩者之間的不同在於 s-結構裡各種非詞組性的訊息也已認明。黃正德教授是漢語語言學界裡的 GB 權威，在他所參與的一個 GB 剖析器研究計劃中，他所採用的模式和 Correa 的有些類似，虛類（empty category）在深層結構中產生，因此也避免了變換律所造成的困擾（Lin et al 1986）。

由此可見，同樣是在 GB 的語法理論架構之下，Chomsky 所主張的形式組在計算機剖析句子的應用上所造成的困難，由於 Correa 和黃正德教授所使用的形式組之不同而得以

避免。

LFG 全名為詞匯功能語法，詞匯（lexicon）佔有重要的地位；句子的詞組結構由詞組結構律所約束，句子的剖析因此相當直接。史丹福大學 CSLI 所設計的 LFG 剖析器和美國 ECS 公司的 LFG 剖析器是很好的例子；兩者之間重要的不同是前者在建構了一個完整的詞組之後才經由聯併建構相對的功能結構，但後者採用了 vLFG 的形式組，在建構詞組任何一部份的同時聯併發生建構相對的部份功能結構。後者的優點在於提早發現不合法的結構及錯誤的剖析路線，因而達到更有效率的剖析過程。

LFG 的功能結構和聯併的運作也是其有效剖析的重要因素。在許多情形下，GB 或類似語法必須產生多種剖析結果，但在 LFG 或其他聯併語法之下則允許單一剖析結果。以下面三個句子為例：

1.a. The deer swims.

b. The deer swim.

c. The deer swam.

在 1a 和 1b 的結構裡動詞或主詞之一標明了是單數或是複數，但是在 1c 裡卻無法得知。因此，在一個非聯併的形式組之下 1c 將造成兩種剖析，主詞分別為單數及複數。在 LFG 的聯併形式組之下卻可以成立單一的樹形及單一的剖析；換言之 deer 在詞匯中亦是單一詞條。

2.deer: CAT N

$$\begin{bmatrix} \text{PRED} & \text{'deer'} \\ \text{NUMBER} & \text{OPT} \end{bmatrix}$$

當然，我們必須公平地指出，從心理語言的角度來看，像1c這樣的句子應該只有一個剖析結果還是有兩種結果，這是一個可實証的（empirical）的問題；其實1c若是在一個自然的語境之下只會有一個分析，因為 deer 的單複數在語境中可以取得，而 LFG 的聯併形式也相容於語境中取得的訊息，因此有效率的剖析是其優點。

四、句子生成

句子生成的過程常是將一個知識表現或是語意結構經由某種語言的規律轉換成句法結構，生成此語言中一個完整的句子以表達這個語意或知識表現的訊息。

GB 式的變換語在生成上遠比在剖析上也有效率，但是 GB 在生成系統上的應用也遠不及 LFG 及其他聯併語法。GB 語法生成句子需以樹形結構為基礎，而 LFG 可以功能結構為開始；前者因語言之不同而有極大的差異，而後者的共通性極強。因此，當由語意或知識表現轉換成句法結構時，LFG 由於有共通性的功能結構的媒介，在程序上比較模組化（modular），因此比較不複雜，但 GB 要從語意及知識直接轉換成各個語言所特有的詞組結構則牽涉較複雜的操作及程式編寫。Grishman（1986:160-168）就描述了一個生成系統，以邏輯表現為基礎，轉換為深層結構，然後再轉換成句子的表層結構。在LFG 理論內部已有關於語意結構和功能結構之間關係的研究，如 Halversen（1983），在史丹福大學所發展的 LFG 語法工作站裡，詞組結構、功能結構、及語意結構是並存的；因此，同樣的過程在 LFG 之下，其程序的模組化及各

模組之間的介面都比在 GB 模式下要自然而有效率。

LFG 的功能結構和聯併運作在句子生成上有另外一個優點，就是並非所有的語法訊息都要完全存在才能生成完整的句子。例如假設有一個提供軍事訊息的系統，在回答「前方接近物體是什麼？」問題時所應該提 供的正確回答是「三艘軍艦」；在生成此正確回答時此一系統由資料庫中先認出接近物體為軍艦，數量為三，將此一訊息化為語意結構再轉為以下的功能結構：

4.[ADJ [FORM 'san 1']

Form 'jun1jian4']

在這功能結構裡唯一的訊息是「軍艦」及數量「三」；沒有量詞無法生成理想完整句子。但是在經過中文生成語法的應用後，這個初期的功能結構裡的詞匯可以得到另外的中文裡特有的訊息。

5.jun1jian4:

[CAT N FS

[Form 'jun1jian4'

CLASS 'sao1']]

6.

[Form 'jun1jian4'

ADJ [Form 'san1']

CLASS 'sao1']

在中文的「軍艦」詞條 5 與功能結構4聯併之後，我們得到的是以上的功能結構6；因為詞組結構的限制，此一功能結構所對應的詞組結構得以生成完整的句子：「三艘軍艦」。

五、機器翻譯

在現有各項機器翻譯的研究計劃中，LFG 極有可能是應用最多的一個語法理論。原因不外是它的理論將詞組結構和功能結構的模組化以及其形式組和聯併的便於運算。反觀 GB 在同一方面的應用就數量而言是遠遠不及的。我們從機器翻譯現有的兩種方式來探究其原因。一種方式是「中間語」（iterlingua），亦是將被譯語（source language）分析之後，將分析所得的結構轉換成一個所有語言共通的語意或知識表現，稱之為「中間語」，然後將中間語的表現轉換為目標語（target language）的句法結構，繼而生成翻譯完整的句子。在這樣的模式之下，如果使用 GB 變換語法，詞組結構要轉換成中間語的過程複雜，因為被翻譯語的詞組結構含有大量的該語言特有的一些現象，同樣的，在將中間語表現直接轉換成目標語的詞組結構也是過於直接而造成繁複的轉換過程。但是以 LFG 為模式則與中間語的雙向介面皆是較具共通性的功能結構；如此較具模組化的過程利於程式編寫的簡易及維護。在著名的 Carnegie Melon 大學的以中間語模式為機器翻譯研究發展計劃 KBMT，其翻譯過程正是以 LFG 為理論基礎。

第二種也是比較實用的一種方式是轉換（transfer），也就是將被翻譯語的分析結果直接做形式上的修改使之成為理想的目標語句子所需要的結構形式，因而由此轉換的過程達成翻譯的目的。同樣的，試想 GB 分析句子所得的被翻譯語的樹形結構要經過如何繁複的修改才能成為目標語所要求的形式，其結構及各語言特有的詞序都需大幅修改轉換；然而，LFG 的

功能結構由於已經脫離了詞組結構及詞序的約束，有相當程度的語言共通性，因此在從被翻譯語的功能結構轉換成目標語的功能結構則是相對地要容易許多。這種轉換上的效率在翻譯兩個愈是形態不同的語言愈是明顯，因為它們的詞組結構和詞序大不相同，但是其功能結構卻仍然是十分相似。

7.a. Taroo ga eigo o hanasu
Taroo GA English o speak

b. Taroo speaks English

7 a-f

```
SUBJ   [FORM 'Taroo'
        PCASE  'ga']
OBJ    [FORM 'eigo'
        PCASE 'o']
PRED   <SUBJ, OBJ>
FORM   'hanasu'
TENSE  PRES
```

7 b-f

```
[SUBJ  [FORM 'Taroo'
        NUMBER SG]
 OBJ   [FORM 'English'
        NUMBER SG]
 PRED  <SUBJ, OBJ>
 FORM  'speak'
 NUMBER SG
 TENSE PRES
```

從7a 可看出日文是賓語在動詞之前，主語賓語均有詞格標記（ case marker ），而7b中的英語卻是主動賓的詞序而且不具有任何詞匯上的詞格標記，因此詞組結構極端不同，但是功能結構卻是十分相似，如7a-f及7b-f所顯示。由此可見，若以 LFG 為基礎，轉換為翻譯模式，其轉換的部份比起以詞組結構為主的語法理論，要簡便許多。

六、結論

自然語言處理研究的核心必須以語言理論為基礎，而在形式組的選擇上也有關鍵性的重要，計算運作的簡便及效率要依靠合適的語言理論及形式組，缺一不可。我們從剖析、生成、及翻譯三方面比較了變換語法 GB 和將語法表現分為詞組及功能結構的 LFG ，表現了後者較前者適於 NLP 的研究，更進一步的顯示了 vLFG 形式組因為同時建立詞組及功能結構，因此在許多方便比起傳統的 LFG 形式組更有效率，更適於 NLP 上的應用。

參考書目

Bresnan, J. 1982 (Ed.). *The Mental Representation of Grammatical Relatons.* Cambridge, Mass: MIT Press.

Bresnan, J. and J. Kanerva. 1989. Locative Inversion in Chichewa: A Case Study of Factorization in Grammar. *Linguistic Inquiry* 20.1:1-50. Also appeared as CSLI Report No. CSLI-88-131, Center for the Study of Language and Information, Stanford University, Standford, CA.

Carbonell J. and M. Tomita. (1987). Knowledge-based Machine Translation, the CMU Approach. In R. Nirenburg, (Ed.). 1987. 68-89.

Chomsky, No. 1957. *Syntactic Structures.* The Hague: Mouton.

Chomsky, N. 1981. Lectures on Government and Binding. Dordrecht: Foris.

Chomsky, N. 1982. Some Concepts and Consequences of the Theory of Government and Binding. Cambridge, Mass: MIT Press.

Chomsky, N. 1986. Knowledge of Language. New York, NY: Preager Publishers.

Correa, N. 1987. An Attribute-Grammar Implementation of Government-binding Theory. Proceedings of 25th Annual Meeting of the ACL.

Crain, S. and J. Fodor. 1985. How Can Grammars Help Parsers? In D. Dowty, et al (Eds.) 1985:95-128.

Gebruers, R. 1989. Book Review: From Syntax to Semantics: Insights from Machine Translation. Machine Translation 4.3:231-238.

Grishman, R. 1986. Computational Linguistics: An Introduction. Cambridge University Press.

Grosz, B., K. Jones, and B. Webber. 1986. (Ed.). Readings in Natural Language Processing. Los Angeles: Morgan Kaufmann Publishers, Inc.

Halvorsen, P.-K. 1983. Semantics for Lexical Functional Grammar. Linguistic Inquiry 14:567-615.

Harris, M. 1985. Introduction to Natural Language Processing. Reston, Virginia: Reston Publishing Company.

Her, O. 1987. The ECS Machine Translation System: an Overview. Technical report and manual, ESC, Inc., Provo, Utah.

Her, O. 1989. An LFG-Based English-Chinese Machine Translation System. Proceedings 1989 International Symposium on Chinese Text Processing, March 16-17, 1989, Boca Raton, Florida. 8.3-8.7.

Her, O. 1989a. Chinese Verb Subcategorization in a Variant Lexical-Functional Grammar. Paper presented at the 22nd International Conference on Sino-Tibetan Languages and Linguistics, October 5-8,1989, Honolulu, Hawaii.

Her, O. 1990. Grammatical Functions and Verb Subcategorization in Mandarin Chinese. Doctoral dissertation, University of Hawaii.

Horrocks, G. 1987. Generative Grammar. London & New York: Longman.

Kaplan, R. 1989. The Formal Architecture of Lexical-Functional Grammar. In Proceedings of ROCLING 11. 1-18.

Kaplan, R. and J. Bresnan. 1982, Lexical-Functional Grammar: A Formal System for Grammatical Representa-

tion. In J. Bresnan (Ed.), The Mental Representation of Grammatical Relations. Cambridge, Mass.: MIT Press. 173-281.

Kaplan, R. and A. Zaenen. 1989a. Long Distance Dependencies, Constituent Structure, and Functional Uncertainty. In M. Baltin & A. Kroch (Eds.), Alternative Conceptions of Phrase Structure. Chicago: University of Chicago Press. 17-42.

Kay, M. 1986. Parsing in Functional Unification Grammar. In Grosz, J. et al 1986. (Ed.).

King, M. 1983. Parsing Natural Language. London: Academic Press.

Kudo, I. and H. Nomura. 1986. Lexical-Functional Transfer: A Transfer Framework in a Machine Translation System based on LFG. Proceedings of Coling 1986, Bonn. 112-114.

Levin, L. 1987. Toward a Linking Theory of Relation Changing Rules in LFG. CSLI Report No. CSLI-87-115, Center for the Study of Language and Information, Stanford University, Stanford, CA.

Lin, L., J. Huang, K. Chen and L. Lee. 1986. A Chinese Natural Language Processing System Based upon the Theory of Empty Categories. Proceedings AAAI 1986 Conference on Artificial Intelligene, Volume II. 1059-1062.

Nirenburg, S. (Ed.). 1987. Machine Translation: Theoretical and Methodological Issues. Cambridge University Press.

Riemsdijk, H. and E. Williams 1986. Introduction to the Theory of Grammar. Cambridge, Mass: MIT Press.

Sells, P. 1985. Lectures on Contemporary Syntatic Theories. Stanford, CA: CSLI, Stanford University.

Shieber, S. 1986. Introduction to Unification-based Approaches to Grammar. Stanford, CA: CSLI, Stanford Uniersity.

Shieber, S. 1987. Separating Linguistic Analyses from Linguistic Theories. In P. Whitelock, et al (Eds.). 1-36.

Slocum, J. 1985. A Survey of Machine Translation: Its History, Current Status, and Future Prospects. Computational Linguistics 11:1-17.

Wedekind, J. 1986. A Concept of Derivation for LFG. Proceedings COLING 1986. 487-489.

Wescoat, M. 1987. Practical Instructions for Working with the Formalism of Lexical Functional Grammar. In J. Bresnan (Ed.). Lexical-Functional Grammar. Course Material for LI229, 1987 Linguistic Institute, Stanford University.

Whitelock, P., M. Wood, H. Somers, R. Johnson, and P. Bennet. (Eds.) 1987. Lingiustic Theory and Computer Application. London: Academic Press.

Winograd, T. 1983. Language as a Cognititve Process (Volume 1: Syntax). Addison-Wesley Publishing Company.

Zaenen, A. and J. Maling. 1983. Passive and Oblique Case. In L. Levin, et al. (Eds.). 159-191.