A Computational Syntax and Its Application to Parsing

Hsin-I Hsieh

Department of East Asian Languages & Literatures
University of Hawaii
Honolulu, Hawaii, U.S.A. 96822
e-mail: hhsieh@uhunix.uhcc.hawaii.edu

Abstract

We propose a theory of syntax in which grammatical sentences are generated by binary composition from lexical frames and syntactic types are computed from lexical types. This theory of computational syntax has an immediate application to the parsing of grammatical sentences. We discuss in detail our theory and briefly demonstrate its applicability to parsing, using Mandarin sentences as examples.

0. Introduction

We propose a theory of syntax in which grammatical sentences are generated by repeatedly composing two elements, each of which belongs to a specific type. At every stage of the binary composition, the type of an element is either specified in the lexicon or is computed by rules based on the types of its two co-composing elements. Thus the syntax is not merely compositional but computational. Grammaticality is precisely characterized: a sentence is grammatical if it belongs to a specific type and also fulfills additional grammaticality conditions.

Our theory of syntax can be applied to the parsing of sentences, especially grammatical sentences. The linguistic part of a parsing system using our syntactic theory can be relatively uncomplicated owing to the internal computation of syntax.

The bulk of this paper is an explication of this computational syntax. We only briefly discuss the applicability of our theory to parsing without actually proposing a full parsing system.

1. Compositional Cognitive Grammar

Our computational syntax is part of a more comprehensive theory of grammar called Compositional Cognitive Grammar (CCG) (Hsieh 1992b). This grammar 'starts' with a component called Imagery Structure (IS) whose elements are called Imagery Structure representations (ISrr). IS maps onto Semantic Structure (SS) whose elements are called Semantic Structure representations (SSrr). SS maps onto Thematic Structure (TS) which contains as its elements Thematic Structure representations (TSrr). TS maps onto Functional Structure (FS) which contains as its members Functional Structure representations (FSrr). FS maps onto Constituent Structure (CS) whose members are Constituent Structure representations (CSrr). Thus the chain of interconnecting mappings is: IS --> SS --> TS --> FS --> CS. At the present stage of our research we are only able to fully articulate the SS and CS components. The TS and FS are lacking but would be similar in purpose to those proposed in LFG (Bresnan 1982). The IS is necessary if the meaning of a sentence is ultimately rooted in its related image, as the cognitive grammarians (Langacker 1987,

Talmy 1985, Jackendoff 1990, Tai 1985, 1989) have suggested or implied. SS is the component where syntax and semantics enter into a systematic interface or interaction (Hsieh 1992a, Chang 1991, Her 1991, M. Hsieh 1992). CS is similar to standard phrase structure. Lacking TS and FS, we map SS onto CS directly. We generate an SSr by repeatedly composing two elements, each one of which is either (i) drawn from a finite set of basic elements listed in the lexicon, or (ii) is a 'persistent' binary composition whose ultimate composing elements are all basic elements. In this sense an SSr is compositional. Each element, basic or non-basic, has a specific pattern or type. If an element is basic, its type is determined in the lexicon, and if non-basic its type is determined by a set of rules of computation applied to its two co-composing elements. In this sense SSr is computational. Hence, as a theory of syntax, SS is both compositional and computational.

1.1. The Lexicon

A sentence or clause in conventional grammar is represented as an Action (AC) in its SSr. An AC may be either a simple Action (sim-AC) or a complex Action (com-AC). Simple Actions compose in a binary way into complex Actions of increasing complexity.

Those simple Actions that are 'legitimate' or well-formed are called Action Frames (ACFs). Each ACF is composed of an Initiator (I) and a complex Act (A'), which in turn is composed of an Act (A) and a Receiver (R). The A is represented by an Abstract Verb (AV) and the I and R may be represented by an indexed variable v_k or a simple constant c or a complex constant f(c), or by an empty symbol 0, which may be indexed by k to express a syntactically empty but pragmatically inferrable NP_k. We say that an ACF 'accepts' a particular AV and an AV 'selects' a particular ACF. An ACF with a particular AV is a particular ACF, or, PACF. Thus an ACF is a type and its associated PACFs are its tokens. For convenience, we sometimes disregard this distinction in our discussion unless it is necessary to maintain it.

The lexicon contains two types of entries: logical-type entries and grammatical-type entries. Logical(-type) entries are divided into three sub-types: (i) indexed variables v_k's; (ii) simple constants c's; (iii) complex constants f(c)'s; (iv) empty symbols without indices, 0's, or empty symbols with indices, 0_k's. An indexed variable v_k is a place-holder for a co-indexed NP_k, which, when suitably attached to the SSr tree, will 'instantiate' v_k by replacing all copies of it. The instantiation operation is based primarily on the concept of instantiation proposed by McCawley (1971) and the notion of controlled empty categories suggested by Huang (1992). A simple constant c indicates the actual site of embedding for an exterior AC indexed by c. A complex constant f(c) refers to a special semantic feature of the AC indexed by c, such as 'aspect', 'tense', 'degree', 'manner', etc. A simple constant c and a complex constant f(c) differ in meaning but behave in the same way syntactically. For convenience, we sometimes write 'c' for both c and f(c) whenever it is convenient to do so. A constant c or f(c) in an ACF is 'saturated' if there is an exterior AC indexed by c which composes with this ACF to supply the content of c. Otherwise, the constant is 'unsaturated'.

Grammatical entries in the lexicon are either abstract verbs (AVs) or nouns (Ns). AVs are divided into three subtypes: (i) full verb (FV); (ii) half verb (HV); and (iii)

grammatical verb (GV). A word (more precisely a morpheme) may function as one or more of these three subtypes. Each AV has a concrete shape in CSr. Roughly, a full verb corresponds to a verb or an adjective; a half verb results in a preposition, conjunction, adverb, auxiliary, tense (marker), aspect (marker), negation (word); and a grammatical verb ends up as a demonstrative, determiner, or a grammatical particle such as the infinitive particle to or gerund affix -ing in English. Exactly which one of these various concrete forms will an AV produce is determined generally by the ACF which the AV selects and occasionally by considering both this ACF and the crucial ACF in its co-composing AC.

A noun may serve as an NP_k to immediately instantiate a co-indexed variable v_k, or may combine with an NP-generating ACF to become the 'core' of an NP_k that instantiates v_k.

ACFs may be classified in two complementary ways: by considering what kind of A (Act) they contain and what kind of I (Initiator) and R (Receiver) they possess. The former consideration yields three v-types and the latter three n-types. The three v-types are: (i) Full Verb AC (FAC); (ii) Half Verb AC (HAC); (iii) Grammatical Verb AC (GAC). An FAC has a full verb for its Act, an HAC a half verb, and a GAC a grammatical verb. The three n-types are: (i) Solitary AC (SAC); (ii) Receptive AC (RAC); and (iii) Warm AC (WAC). An SAC has no unsaturated constants; an RAC contains only one unsaturated constant; and a WAC possesses two unsaturated constants.

As shown in Table I, there are a total of 27 (twenty-seven) ACFs which fall into 9 (nine) compound types. Within each type, the ACFs are further distinguished

Table I. Action Frames Classified.

	<I, <A, R>>			<I, <A, R>>			<I, <A, R>>		
	(1)	x_i	FV y_j	(10)	x_i	HV y_j	(19)	x_i	GV y_j
	(2)	x_i	FV 0	(11)	x_i	HV 0	(20)	x_i	GV 0
SAC	(3)	0	FV y_j	(12)	0	HV y_j	(21)	0	GV y_j
	(4)	0	FV 0	(13)	0	HV 0	(22)	0	GV 0
	(5)	h	FV y_j	(14)	h	HV y_j	(23)	h	GV y_j
	(6)	h	FV 0	(15)	h	HV 0	(24)	h	GV 0
RAC	(7)	x_i	FV k	(16)	x_i	HV k	(25)	x_i	GV k
	(8)	0	FV k	(17)	0	HV k	(26)	0	GV k
WAC	(9)	h	FV k	(18)	h	HV k	(27)	h	GV k
		FAC			HAC			GAC	

FV = Full Verb (verb, adjective).
HV = Half Verb (preposition, conjunction, adverb, auxiliary, aspect, tense, negation).
GV = Grammatical Verb (demonstrative, determiner and grammatical particles).

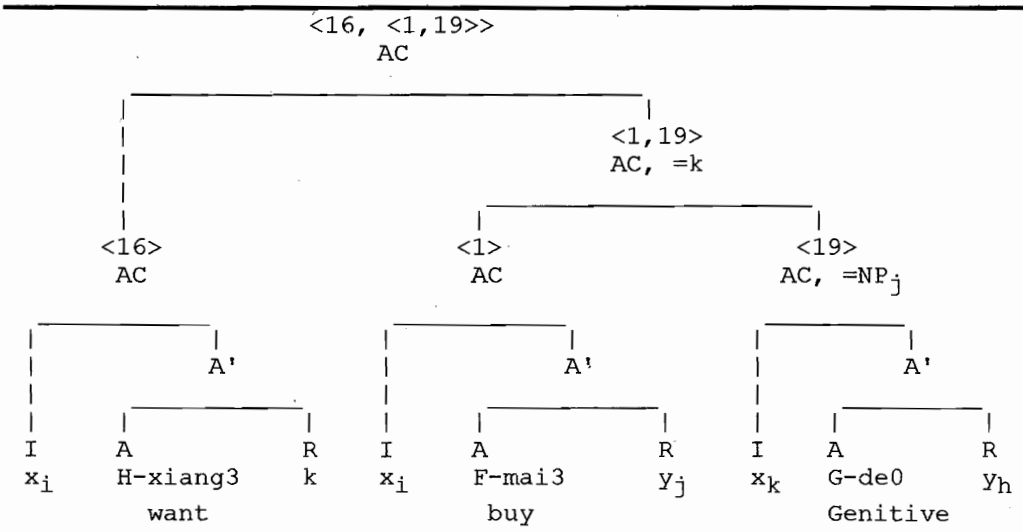
SAC = Solitary AC, containing no unsaturated constants.
RAC = Receptive AC, containing one unsaturated constant.
WAC = Warm AC, containing two unsaturated constants.

k = k or f(k); h = h or f(h).

according to the kind of logical symbols that represent their I and R, namely, v_k, 0, c. For convenience, a v_k is written x_k if it represents I and y_k if it represents R. Also for convenience, a constant c is written h if it represents I and k if it represents R.

A lexical entry is then just a variable v_k, a constant c or f(c), an empty symbol 0, an N, or an AV. This treatment is a technical realization of an insight originated with James H-Y. Tai, who has on many occasions told me that we need to recognize only nouns and verbs in the 'deep' structure of a Chinese sentence. If an entry is an AV, it will select one or more of the 27 ACFs. Each selected ACF, sometimes together with a co-occurring ACF, will determine the concrete form of this AV in that ACF. Each AV in an ACF thus has an abstract category as full, half, or grammatical verb, and a matching concrete category as verb, preposition, demonstrative, etc.

Figure 1 illustrates the composition of ACFs into a com-AC. There we see that ACF <1> and ACF <19> combine to form a com-AC denoted as AC <1,19>, which in turn combines with ACF <16> to form a com-AC denoted as AC <16, <1,19>>. AC <19> is an NP-generating ACF and here it has the effect of creating the NP_j, Li3-si4 de0 che1, which instantiates the y_j in ACF <1>. AC <1,19>, marked as equal to the constant k by the sign '=k', saturates the constant k, which lexicalizes the R in ACF <16>. Within each ACF, the AV is prefixed by F-, H-, or G- to indicate its status in terms of being full, half, or grammatical. In the CSr the variables x_i, y_j, x_k and y_h will be properly instantiated by co-indexed NPs, the AVs given their concrete forms, and the tree will be compressed into a conventional phrase structure by a host of transformations. The result of these operations will yield the CSr for sentence (1).



(1) Zhang1-san1 xiang3 mai3 Li3-si4 de0 che1.
'Zhang-san wants to buy Li-si's car.'

(Note: The prefix F-, H-, or G- to an AV indicates the full, half, or grammatical status of the AV.)

Figure 1. (partial) SSr for (1).

AC <16, <1,19>> illustrates the notion that in the SSr every com-AC is ultimately composed of a finite number of ACFs. In other words, every sentence is compositionally derived and not 'generatively' derived via a set of rewriting rules whose initial symbol is S, as is practiced in most current theories including GB.

1.2. Computation

Although we are free to compose ACFs of various kinds to form a complex AC, not all ACs so formed will have a CSr that yields a grammatical sentence. In other words, only some ACs are well-formed. There are syntactic, semantic, and pragmatic types of well-formedness conditions for an AC. An AC satisfying all these types of well-formedness conditions would be a 'maximally well- formed' AC. Such maximally well-formed ACs are beyond the scope of our discussion, since our focus here is on syntax. Nevertheless, we can identify a syntactic well-formedness condition, based on compound types, which a 'minimally well-formed AC' must meet. A grammatical sentence in a language that requires a compulsory marking for both aspect and tense can be analyzed into three parts: a 'core' content, an aspect, and a tense. To translate this arrangement into a composition in terms of ACs, we can first combine the core content with the aspect, and then combine the result with the tense. Suppose that the core is of the type FAC-SAC, and the aspect is of the type HAC-RAC. Then the composition of the core and the aspect could have the type FAC-SAC. Now suppose that the tense is also of the type HAC-RAC, then composing the core-aspect combination with tense would yield also the type FAC-SAC, to which all grammatical sentences could belong. Thus we have a clear idea of what the syntactic well-formedness condition for a minimally well-formed AC would be. It would be this: every minimally well- formed AC must have the type structure stated in (wf): <<FAC-SAC (core), HAC- RAC (aspect)>, HAC-RAC (tense)>. For example, sentence (2) John has seen Mary will have the SSr in (2'): $<<<x_i$, <F-see, $y_j>>$ =k, <aspect (k), <H-perfect, 0>>> =h, <tense (h), <H-present, 0>>>. The 'compound-type structure' of (2') is (2'(a)) <<FAC-SAC, HAC-RAC>, HAC-RAC> and its 'compositional structure' in terms of ACFs, or, 'ACFs structure' is (2'(b)) <<1,15>, 15>. (2'(a)) as a compound-type structure can be computed to yield the type FAC-SAC, and for this reason it is the computational part of (2'). (2'(b)) as an ACFs structure is a composition not subject to any computation and so it is the compositional part of (2'). We combine (2'(a)) and (2'(b)) into (2'(ab)) <<FAC-SAC (1), HAC-RAC (15)>, HAC-RAC (15)>, and we obtain the 'compositional- computational complex' (ccc) of (2') and eventually of (2).

Clearly, we want to set up our computational scheme in such a way as to ensure that every grammatical sentence will end up having the (wf) as its compound-type structure. To achieve this, we make sure that when FAC-SAC is combined with a regular compound-type, it will yield FAC-SAC, but when combined with an exceptional type, it will yield the same exceptional type. In other words, we need two guiding principles: (i) FAC-SAC combined with x will yield FAC-SAC, and (ii) FAC-SAC combined with y will yield y. Apart from ensuring that every grammatical sentence has the (wf) as its type structure, we also want to prohibit incompatible ACFs from entering into composition. Specifically, (iii) we need to rule out the composition of two RACs or two WACs or one RAC with one WAC. By considering all these three needs, we decide that the computation of v-types should follow rules as

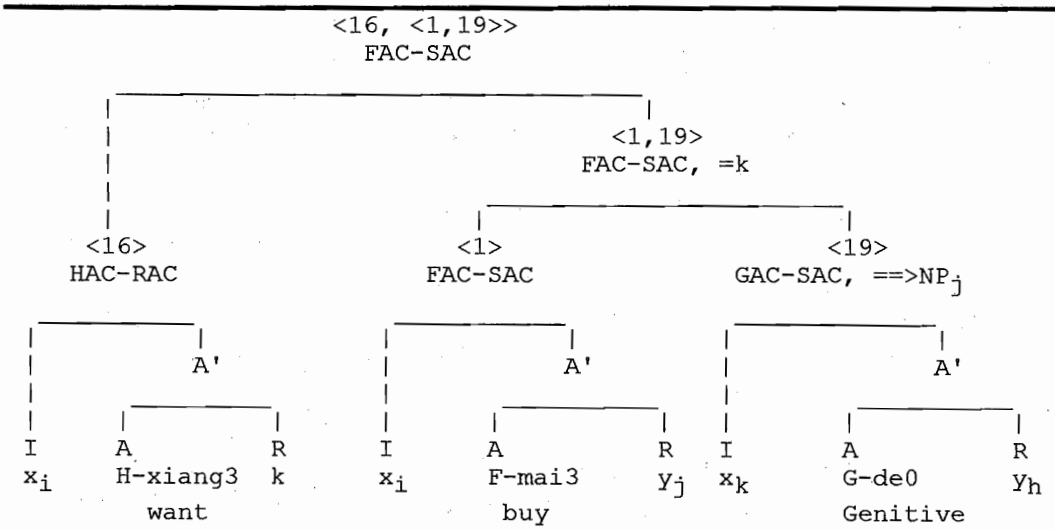
Table II. Computation of v-types.

	FAC	HAC	GAC
FAC	(i) FAC	(ii) FAC	(iii) (a) FAC, if GAC is an SAC (b) GAC, if GAC is an RAC or WAC
HAC		(iv) HAC	(v) (a) HAC, if GAC is an SAC (b) GAC, if GAC is an RAC or WAC
GAC			(vi) GAC

Table III. Computation of n-types.

	SAC	RAC	WAC
SAC	(i) SAC	(ii) SAC	(iii) RAC
RAC		(iv) Goof	(v) Goof
WAC			(vi) Goof

Goof: A complex AC which is not allowed to combine with any AC of the type SAC, RAC, WAC.



(1) Zhang1-san1 xiang3 mai3 Li3-si4 de0 che1.
'Zhang-san wants/intends to buy Li-si's car.'

Figure 2. Computation of Compound Types for (1).

stipulated in Table II and the computation of n-types should obey rules as prescribed in Table III. These two tables are self-explanatory and require no comments except that the computation is commutative as is obvious from the fact that only the upper- right halves of the tables are shown. The two tables work jointly to assign to a composition of any degree of complexity its combination of v-type and n- type, that is, its compound type.

In Figure 2 we illustrate this operation by showing how to derive the compound type for the SSr of sentence (1). Here we see that FAC in (ACF) <1> combines with GAC in <19> to become FAC in <1,19>, following rule (iii)(a) (since the GAC is also an SAC) in Table II. On the other hand, SAC in <1> combines with SAC in <19> to become SAC in <1,19>, following rule (i) in Table III. Combining these two results, we obtain the compound type FAC-SAC for <1,19>. Subsequently, the HAC of <16> combines with the FAC of <1,19> to form FAC, according to rule (ii) in Table II. Simultaneously, RAC of <16> combines with SAC of <1,19> to form SAC, according to rule (ii) in Table III. Putting these two results together, we obtain FAC-SAC for <16, <1,19>>.

1.3. Word Order

The SSr for sentence (1) shown in Figure 2 is explicit about the hierarchical order of the elements (of various complexity) in composition but it still needs information that would determine the linear order of these elements. Linear orders are determined by the 'primacy relation' which holds between a 'primary' ('p') ranked element and a 'secondary' ('s') ranked element in a composition. Each element alone has a 'p' rank, but when two elements are composed, one of them retains its 'p' rank and the other is demoted to an 's' rank. Within an ACF the primacy relation is predetermined and has the configuration AC(p) = <I(s), <A(p), R(s)> = A'(p)>. Across two ACs, the 'p' and 's' ranks are assigned by comparing their degrees of primacy. If one has a higher degree than the other, it is ranked 'p' and the other is ranked 's'. If both have the same degrees of primacy, then semantic and other non-syntactic criteria apply to determine their 'p' and 's' ranks.

The generally distinct degrees of primacy for a pair of co-composing elements are obtained by combining the result of computation based on the v- types with that based on the n-types. The laws of computation for the v-types are set forth in Table IV and those for the n-types are stated in Table V. These two tables are self-explanatory. Two compound types under comparison may be represented as the ordered pair $<v_1n_1, v_2n_2>$ (e.g. <FAC-SAC, GAC-SAC>). The v-type computation will yield an ordered pair $<d(v_1), d(v_2)>$, with $d(v_1)$ and $d(v_2)$ representing the degrees of primacy for v_1 and v_2, respectively. Similarly, the n-type computation will produce an ordered pair $<d(n_1), d(n_2)>$, with $d(n_1)$ and $d(n_2)$ denoting the degrees of primacy for n_1 and n_2, respectively. We then compute the sum of the two ordered pairs by adding up their coordinates and we obtain the sum $<d(v_1) + d(n_1), d(v_2) + d(n_2)>$, which is simply the ordered pair <m,n>, with m and n being positive integers. This pair <m,n> indicates the degrees of primacy for $<v_1n_1, v_2n_2>$, the two compound types under comparison. In other words, $<m,n> = <d(v_1n_1), d(v_2n_2)>$, where $d(v_1n_1)$ is the primacy degree of type v_1n_1, and $d(v_2n_2)$ is the primacy degree of type v_2n_2. If $m > n$, then m is

interpreted as 'p' ('primary') and n as 's' ('secondary'). If m < n, then m is interpreted as 's' ('secondary') and n as 'p' ('primary'). If m = n, then other criteria must apply to interpret one of m and n as 'p' and the other as 's'. Figure 3 illustrates this method of primacy ranking using sentence (1) as example. Here we see that elements within each ACF are assigned their 'p' and 's' through predetermination. Across ACF <1> and ACF <19>, a computation shows that the pair <m,n> representing the primacy degrees of ACF <1> and ACF <19> has the actual value of <2 + 1 = 3, 0 + 1 = 1>, with 3 > 1, hence ACF <1> is ranked 'p' and ACF <19> is ranked 's'. Across ACF <16> and AC <1,19>, the value for <m,n> is <0 + 1 = 1, 2 + 0 = 2>, with 1 < 2, hence ACF <16> is 's' and AC <1,19> is 'p'. Finally, <16, <1,19>> is assigned 'p' since it stands alone. Were it to enter into a composition, its primacy rank may be adjusted.

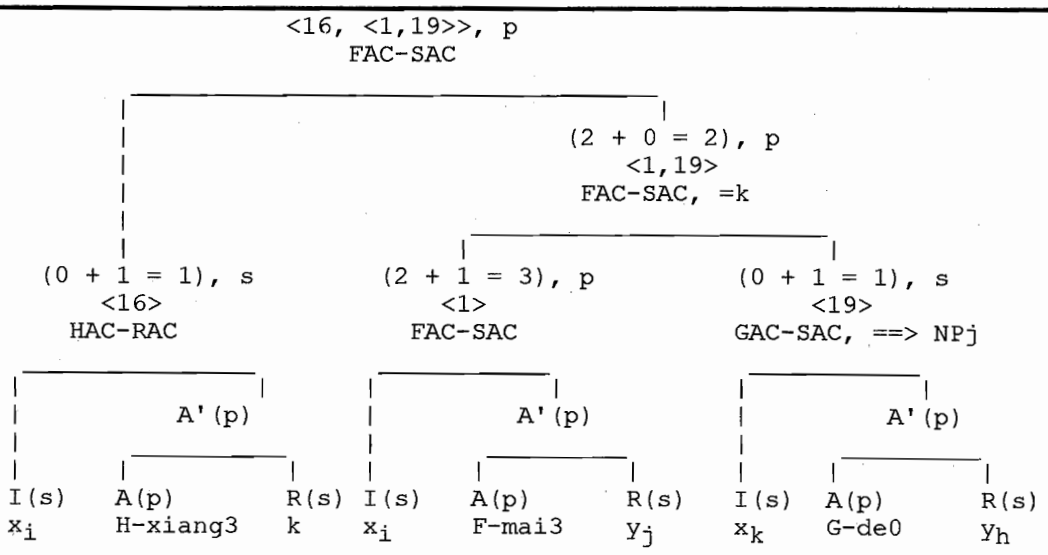
Once we have obtained the 'p' and 's' ranks for the pair of composing elements in each composition, we can interpret the 'p'-to-'s' relation in a particular language as

Table IV. Computing the Primacy Degrees in v-types.

	FAC	HAC	GAC
FAC	(i) <2,2>	(ii) <2,0>	(iii) <2,0>
HAC	(iv) <0,2>	(v) <2,2>	(vi) <2,0>
GAC	(vii) <0,2>	(viii) <0,2>	(ix) <2,2>

Table V. Computing the Primacy Degrees in n-types.

	SAC	RAC	WAC
SAC	(i) <1,1>	(ii) <0,1>	(iii) <0,1>
RAC	(iv) <1,0>	(v) none	(vi) none
WAC	(vii) <1,0>	(viii) none	(ix) none



(1) Zhang1-san1 xiang3 mai3 Li3-si4 de0 che1
'Zhang-san wants to buy Li-si's car.'

Figure 3. Computation of Primacy Degrees for (1).

either a 'p'-preceding-'s' or an 's'-preceding-'p' relation in word order. Although the picture is somewhat complicated, in Mandarin Chinese the general rule of word order is an 's'-preceding-'p' order, as Huang (1982, 1993), Li (1985, 1990), and Tai (1973) have shown, using a head versus modifier distinction, which is roughly parallel to our 'p' versus 's' distinction. Notable exceptions include VO order and prepositions.

The SSr for (1) as given in Figure 3 contains all the specifications needed for an SSr and as such it is a 'complete' SSr, ready to turn into a CSr through a series of transformations involving movement, concretization, instantiation, pruning, and rebuilding. We now describe these transformations using as example sentence (3) whose SSr is shown in Figure (4a).

2. Application to Parsing

By making use of the SSr, we can parse a grammatical sentence with relative ease. For each semantic interpretation of a surface grammatical sentence, there is exactly one SSr. This SSr is determined by its ultimate composing ACFs, or, more precisely PACFs. Once we successfully identify these ACFs and their compositional structures, the generation of the final SSr is automatic, since the compound-type formation rules and the primacy-rank determination rules will apply in a computational manner. After the grammatical sentence under parsing is successfully divided into a number of k segments, each containing (the concrete form of) an AV, our task is to search for the ACF that accepts each such AV, and to determine how these ACFs are composed together. Although each AV can select from a small range of ACFs, we can stipulate a set of ACF acceptance rules (ACFARs) which will assist us to decide on a unique correct ACF. We need also a set of ACFs composition rules (ACFCRs) that determine the composition of ACFs based on their compound types and perhaps also on the AVs they accept. The central linguistic problem in parsing with SSrr then is reduced to the formulation of the ACFARs and ACFCRs. These rules are specific to the language under parsing and are finite in number and so in principle can be exhaustively discovered. Thus the potential application of CCG, with its SSrr, to the parsing of grammatical sentences is in principle an uncomplicated and accomplishable job.

References

Bresnan, Joan. (ed.) 1982. The mental representation of grammatical relations. Cambridge, MA: MIT Press.

Chang, Claire Hsun-huei. 1990. Complex verb and argument structure: interaction between syntax and morphology. Paper presented at the Second Northeast Conference on Chinese Linguistics, The University of Pennsylvania, Philadelphia, May 4-6, 1990.

_____. 1991. Verb copying: Towards a balance between formalism and functionalism. Journal of Chinese Language Teachers Association 26.1:1- 32.

Her, One-soon. 1991. Topic as a grammatical function in Chinese. Lingua 84.1-23.

Hsieh, Hsin-I. 1992a. In search of a grammatical foundation for dialect subgrouping. Symposium series of the Institute of History and Philology, Academia Sinica, no. 2: Chinese languages and linguistics, vol. 1: Chinese dialects, 333-377. Taipei: Academia Sinica.

_____. 1992b. Lexicon and morphology in a compositional cognitive grammar. Proceedings of IsCLL-3, 38-61.

Hsieh, Miao-ling. 1992. Analogy as a type of interaction. Journal of Chinese Language Teachers Association 28.3:75-92.

Huang, C.-T. James. 1982. Logical relations in Chinese and the theory of grammar. Doctoral dissertation, MIT.

_____. 1992. Complex predicates in control. Control and grammar, ed. by R. K. Larson, S. Iatridou, U. Lahiri and J. Higginbotham, 109-147. Dodrecht: Kluwer Academic Publishers.

_____. 1993. More on Chinese word order and parametric theory. MS.

Jackendoff, Ray. 1990. Semantic structure. Cambridge, MA: MIT Press.

Langacker, Ronald W. 1987. Foundation of cognitive grammar, vol. 1: Theoretical prerequisites. Stanford: Stanford University Press.

Li, Y.-H. Audrey. 1990. Order and constituency in Mandarin Chinese. Dordrecht: Kluwer Academic Publishers.

McCawley, James. 1971. Where do noun phrases come from? Semantics, ed. by D. D. Steinberg and L. A. Jakobovits, 217-231. Cambridge: Cambridge University Press.

Tai, James H-Y. 1973. Chinese as a SOV language. Papers from the 9th Chicago Linguistic Society meeting, 659-71. Chicago: Chicago Linguistic Society.

_____. 1985. Temporal sequence and Chinese word order. Iconicity in syntax, ed. by John Haiman, 49-72. Amsterdam: John Benjamins Publishing Co.

_____. 1989. Toward a cognition-based functional grammar in Chinese. Functionalism and Chinese grammar, ed. by James H-Y. Tai and Frank F. S. Hsueh, 187-226. Chinese Language Teachers Association Monograph Series No. 1.

Talmy, Leonard. 1985. Lexicalization patterns: semantic structure in lexical forms. Language typology and syntactic description, vol. 3: grammatical categories and the lexicon, ed. by Timothy Shopen, 57-149. Cambridge: Cambridge University Press.

使用新式注音鍵盤及複合馬可夫語言模型之中文輸入系統

A Chinese-character Inputting System Using a New Type of Phonetic Keyboard and a Compound Markov Language Model

古 鴻 炎　　陳 志 耀
Hung-yan Gu and Jr-yiau Chen

國立台灣工業技術學院 電機系
Department of Electrical Engineering
National Taiwan Institute of Technology
Taipei, Taiwan, R. O. C.
E-mail: gu@mouse.ee.ntit.edu.tw

摘要

我們設計、製作了一個實際且方便的中文輸入系統，系統裡採用宜韻鍵盤之新式注音按鍵排列來輸入國語音節之聲、韻、調符號；此外，使用者不需逐音選字，而由系統自動作音至字轉換的處理來找出一個最可能的中文句子，如果仍然有錯字，則可由使用者很方便地加以更正。由於系統的詞典所收錄的詞是有限的，因此，我們製作提供了線上詞彙學習的功能，包括新詞登錄，舊詞出現頻率調整，及舊詞刪除。此外，考慮到中文字裡有許多不常用的字，一般人可能不會唸而無法輸入，因此，我們提出近形字群線上查詢與建立的想法，並加以實現。

關於自動音轉字的問題，我們提出一種複合式馬可夫語言模型來解決，它不但支援線上詞彙學習功能，也將句子裡相鄰兩詞間的相關性考慮進來。對於此模型，我們曾以一篇社論文章及一篇兒童故事來測試其轉換率，初步結果是，所提出的模型比零階、詞爲狀態之馬可夫模型爲好，並且比一階、字爲狀態之馬可夫模型穩定。

關鍵詞：中文輸入、注音鍵盤、馬可夫模型

國科會補助專題研究計劃編號： NSC 82-0408-E011-209

一、導言

使用電腦來處理中文資訊的一個嚴重瓶頸在於中文之輸入，爲了提高電腦在中文社會的普及率，並加速中文資訊之電腦化，我們覺得各種中文輸入方法的發展都是很有意義的。目前可見的中文輸入方法，它們所採用的輸入媒介包括鍵盤、滑鼠〔1〕、感應筆(線上手寫輸入)〔2,3〕、掃描器(光學文字辨識)〔4,5〕、語音(語音辨識輸入)〔6〕等，其中，鍵盤仍是目前在輸入效率、操作省力性、費用等因素同時考慮下，一種較佳的中文輸入媒介，不過，由於中文字的個數是數以萬計的，如目前廣泛被採用的big-5中文內碼〔7〕，有一萬三千左右的字數，所以，我們很難以鍵盤來直接輸入中文，而實際上被採用的作法有：(1)賦予各個中文字一個數字代碼，而以鍵盤來輸入數字代碼；(2)依據一些規則將各個中文字拆解成字根，然後以鍵盤來輸入字根，例如倉頡、大易、行列等輸入法；(3)依據各個中文字之讀音，以鍵盤輸入此音之注音符號，例如漢音〔8〕、國音〔9〕、忘形〔7〕等輸入法。雖然目前已有各式各樣的中文輸入方法，但是，並沒有一種方法能夠說是十全十美的，可能被抱怨的缺點如：需花一段時間學習輸入方法且隔一段時間不用又可能忘記(如字根類之中文輸入方法)；原本之輸入速度就不快，連續使用後會因手或眼疲勞而輸入更慢(如滑鼠、線上手寫之輸入方式)；設備費用較爲昂貴(如光學文字辨識)等。

考慮數種輸入方式的優缺點後，我們決定使用原本就已配備的鍵盤來作爲輸入中文的媒介，如此，可不必多花額外費用，熟練後可盲目打鍵(打鍵時不看鍵盤)較不易讓手、眼疲勞，並且輸入速度之上限比滑鼠與手寫方式高。此外，我們也選擇以中文字之國語讀音作爲間接輸入中文之依據，這是考慮大多數人都有學過國語注音符號，而可省去學習使用的時間，並去除可能的恐懼、疑慮。由於原始之注音輸入法需由使用者逐音選取螢幕上出現的同音字，而使輸入速度快不起來，並且無法盲目打鍵，因此，我們建造的系統將具備自動選取同音字的功能，而選錯時可由使用者方便地加以更正，如此，就可使輸入速度及盲目打鍵問題獲得舒解。

整體來看，我們的系統是由如圖1裡的幾個功能方塊所組成，其中，鍵盤

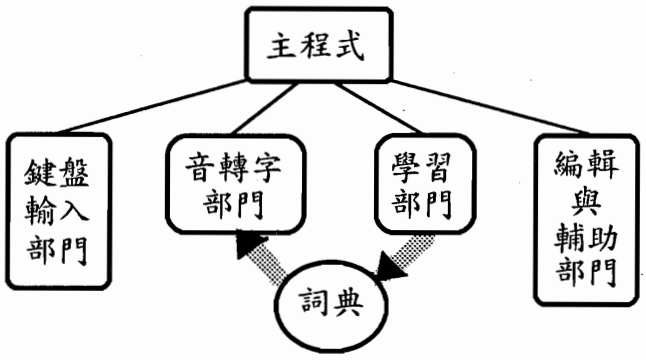


圖1 系統架構

輸入部門負責按鍵之讀取與解釋；音轉字部門，負責將使用者輸入的音節序列轉換成最可能的中文句子；學習部門負責接受使用者的指示去登錄新詞、消掉舊詞、或調整舊詞之出現頻率，此外，它也提供線上近形字群建立的功能；編輯與輔助部門，除了提供文字編輯的功能之外，也能在收到特定按鍵後去呼叫其它輔助功能，如查單一(或全篇)字之注音，查某一字之同音字或近形字等。在以下各節將會對各功能方塊作較詳細之說明。

關於系統之製作，我們採用 Watcom C/386 Ver. 9.01 之 C 語言編譯器，它支援個人電腦內延伸記憶體(extended memory)之使用，不受 640 Kbytes 的限制。當執行所建造的系統後，大約會佔據 1.55 Mbytes 的記憶體(其中 1.15 Mbytes 是延伸記憶體)，此時，螢幕上會顯示出如圖2的畫面，畫面上方的

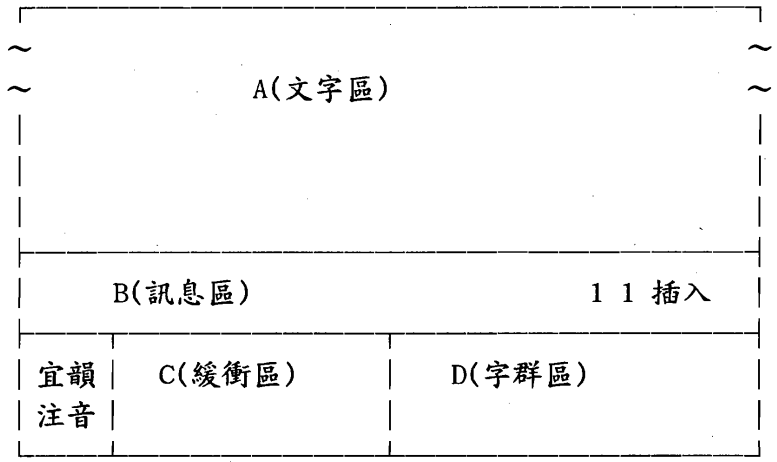


圖2 螢幕畫面

"A"視窗，我們稱它爲文字區，用以顯示已輸入(由檔案或鍵盤)的文字，並且編輯的動作也在本視窗進行；下面的"C"視窗稱爲緩衝區，是使用宜韻注音鍵盤來輸入中文字的視窗，使用者可隨時隨意在"A"、"C"兩視窗之間作切換；另外，"B"視窗是系統用以顯示警告、動作完成等訊息的，而"D"視窗是用以顯示使用者所查詢的同音字、詞，及近形字群等，這是指在緩衝區內所進行的查詢，如果是在文字區進行查詢時，則以機動調整位置之視窗來顯示。所以本系統的使用者介面，因採用具有不同顏色之數個視窗而不會顯得呆板。

二、鍵盤輸入部門

這個部門負責處理使用者由鍵盤輸入的按鍵，它內部有兩個處裡模式，其中一個處理模式(於圖2之文字區時生效)將鍵盤當成是平常之英文鍵盤，對所獲得之輸入碼不另外作解釋，因此，當我們的系統是在進入某一個基礎中文系統(如倚天中文系統)後再執行時，就可利用此模式來輸入英文或憑藉基礎中文

系統所提供的中文輸入方法來輸入中文字。然而，當被切換至另一個處理模式(於圖2之緩衝區)時，此部門就將鍵盤當成是宜韻注音鍵盤〔10〕，並對所獲得之輸入碼加以解釋，以轉換成宜韻鍵盤上對應之注音符號。

在本系統裡實現(寫作驅動軟體)之宜韻注音鍵盤，它是由本文第一作者先前所設計的，同時考慮了三項鍵盤設計的準則，即鍵盤效率，人體工學原則，及符號至按鍵對應的規律性。鍵盤效率指的是輸入一個音節的平均按鍵次數；而人體工學原則是要盡量減少手指頭的運動量，以避免疲勞；至於符鍵對應之規律性，是要讓使用者很輕鬆地建立符號和按鍵位置的聯想對應關係。在這三個準則需同時考慮的條件下，宜韻鍵盤可說是一個相當不錯的設計。

宜韻注音鍵盤把一個實體鍵盤當成兩個虛擬鍵盤來使用，分別稱爲聲母鍵盤與韻調鍵盤，聲母鍵盤如圖3所示，是依一定規則將國語音節裡用到的聲母放上去，而韻調鍵盤如圖4所示，也是依一定規則將國語音節用到的韻母及聲調佈放上去，這裡所說的規則是，在圖3裡，各鍵列上的聲母都是由中間向外邊排列開來，並且左右手兩邊都是由上而下排下來，而在圖4裡，帶介音"ㄧ"的韻母都排在最上列，帶介音"ㄨ"的韻母都排在次上列，帶介音"ㄩ"的韻母都排在主鍵列，並且，各列是按照ㄧ、ㄨ、ㄩ的順序由上而下排下來，其它不帶介音的韻母大多排在最下列，左手四手指由外而內分別負責打以ㄞ、ㄟ、ㄠ、ㄡ爲韻尾的韻母，右手四手指由外而內分別負責打以ㄢ、ㄣ、ㄤ、ㄥ爲韻尾之韻母。前述之排列規則，看似簡單，實際上是經過精心的設計，所以在實行上才不會產生困難，即機器可追蹤使用者按鍵的次序來確定他要輸入那一個注音成分(聲母，韻母，或聲調)，並且使用者不需按特殊鍵去作鍵盤面的切換、或告知系統已輸入一個音節了(即可連續輸入一序列的音節)，至於空韻音節(如/ㄓˇ/)之輸入，所採取的補救辦法是將這類音節的聲調按鍵從主鍵列提到上鍵列，因爲和空韻連接的聲母不會和帶有介音/ㄧ/的韻母連接。如此，輸入一個音節只需按2至3鍵，而由統計分析得知，輸入一個音節平均需按2.76次按鍵〔10〕，比一般注音鍵盤之3.12次按鍵少了0.36次，這意謂每輸入1000個音節可少按360次按鍵，也就節省了相當多的時間及力氣。另外，宜韻鍵盤所遵

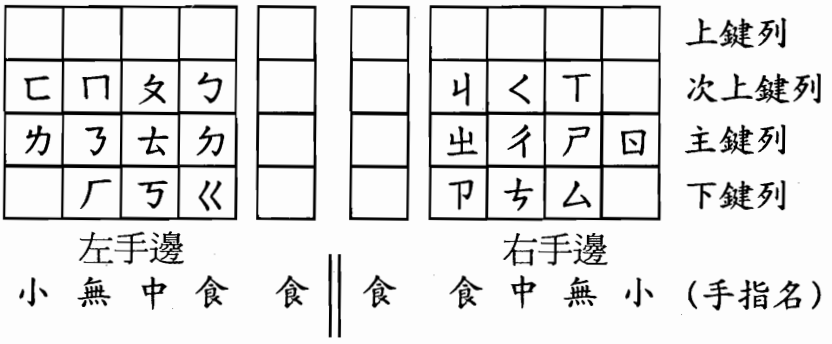


圖3 聲母鍵盤

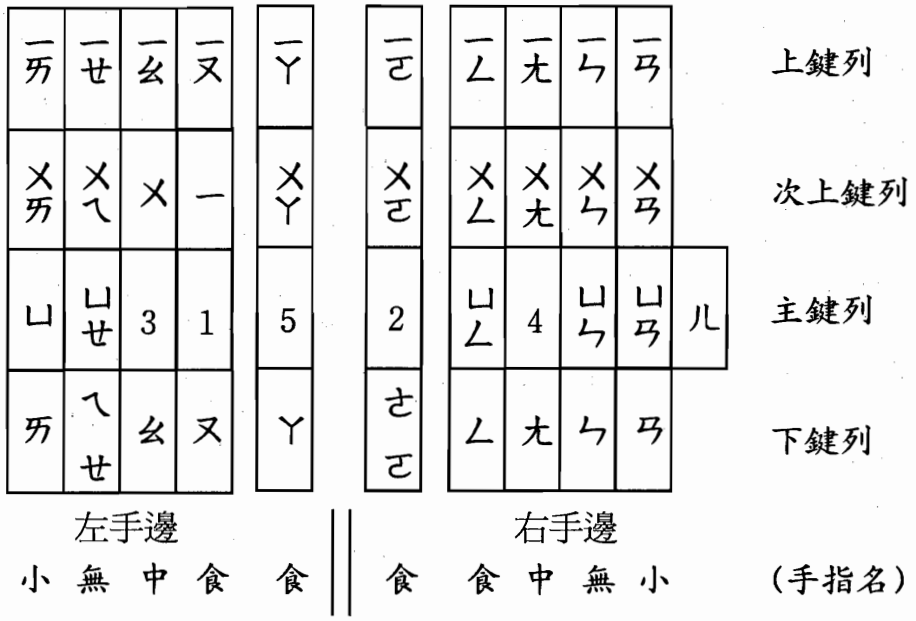


圖4 韻調鍵盤

循的兩個人體工學原則是，各鍵列的打鍵負擔分佈中，主鍵列的百分比應該是最大的，而離主鍵列最遠的上鍵列，它的百分比應該最小；各手指的打鍵負擔分佈中，食指的百分比應該比中指的大，中指的應比無名指的大，而無名指的應比小指的大，這是因爲各手指的靈活程度有差別，其中食指最靈活。

三、音轉字部門

以注音鍵盤來輸入中文的輸入方式，早期需由使用者在輸入每一個音節後，立刻從螢幕上顯示出的同音字中去挑選一個所要的字，這樣，眼睛需交替盯著螢幕(以選字)及鍵盤(以輸入注音符號)，自然使輸入速度快不起來。因此，本系統的音轉字部門提供了自動音轉字的功能，以便讓使用者能夠一口氣把一句話的音節注音整批輸入進去，事實上我們的音轉字部門是採取動態規劃演算法來作漸進的的處理〔11,12〕，如此，使用者每輸入一個音節後，音轉字部門就能夠立刻把目前最可能被對應的中文句子顯示出來，不過，還是希望使用者在輸入一個句子後才去檢查有無挑到非想要的字，然後再加以更正(這樣的機會一般說來低於10%)，如此就可讓輸入速度提高許多。

目前被提出來解決自動音轉字問題的方法，至少有直接查詞典法〔13〕、統計法〔14,15〕、語法剖析法〔16〕等，而我們所提出的方法可算是一種統計作法，其想法是要把詞頻資訊和相鄰字間的相關性資訊加以組合運用，正式說來就是把以詞爲狀態之零階馬可夫模型(簡稱 MW0 模型)和以字爲狀態之一階馬可夫模型(簡稱 MC1 模型)組合成一個複合馬可夫語言模型，希望藉以提高自

動音轉字的正確率。根據我們過去所做的研究〔11,12〕得知，以詞爲狀態之馬可夫模型(零階或一階)，其轉換率比以字爲狀態之馬可夫模型較高且穩定，依據這樣的觀察，原本應選擇採用以詞爲狀態之一階馬可夫模型，但是考慮到我們的系統將提供線上新舊詞學習的功能，其中一項是允許使用者增加新詞到詞典去，需要加入一個新詞是預期它將經常被用到，但此時有關這個新詞與其它舊詞連接時的相關性資料系統裡卻沒有，而使得MW1模型不會選到此新詞，例如增加新詞"漢堡"到詞典去以後，MW1模型對條件機率 P(漢堡|吃) 的估計值仍然是很小(如果訓練語料裡，"吃"與"漢堡"沒有連在一起出現過的話)，因此，我們只得採用MW0模型，以使新增的詞有較大的影響力。接著，爲了使句子裡詞與詞連接的相關性資訊也能用於作音轉字決擇時的參考，且要免除新詞造成的困難，所以，我們就提出了一個變通的作法來估計某二詞相鄰的可能機率，其實就是用前一詞之詞尾字與後一詞之詞頭字相鄰的機率來取代，而這樣的機率值可以MC1模型來估計，例如一個由"他"、"是"、"游泳"、"選手"等詞所串接成的句子，我們提出的複合式模型將以下式

$$P(\text{他是游泳選手}) = P(\text{他}) * P(\text{是}|\text{他})^{W} * P(\text{是}) * P(\text{游}|\text{是})^{W} * P(\text{游泳}) * P(\text{選}|\text{泳})^{W} * P(\text{選手}) \quad (1)$$

來估計此句子之出現機率，其中 w 是加權常數，其數值要由實驗來決定。如果只以MW0模型來估計，則計算式子爲

$$P(\text{他是游泳選手}) = P(\text{他}) * P(\text{是}) * P(\text{游泳}) * P(\text{選手}) \quad (2)$$

另外，如果只以MC1模型來估計，則計算式子爲

$$P(\text{他是游泳選手}) = P(\text{他}) * P(\text{是}|\text{他}) * P(\text{游}|\text{是}) * P(\text{泳}|\text{游}) * P(\text{選}|\text{泳}) * P(\text{手}|\text{選}) \quad (3)$$

前面的式子(2)與(3)是假設中文具有馬可夫特性而推導得到〔11,17〕，至於式子裡的條件機率項 P(Y|X)，並不是要在模型訓練時就將所有可能的 X、Y 組合之 P(Y|X) 值算出來，而事實上也不可能，因爲可能的組合多於 10,000*10,000 ，並且即使收集了非常大量的語料後，仍然會有許多組合不曾看過，這就是所謂的零出現頻率問題，對於這樣的問題已有一些專家提出了不錯的解決方法〔18,19〕，使得式子(1)至(3)在實做上不成問題。對於P(Y|X)之估計值的求取，我們採用的作法如下列式子所示：

$$P(Y|X) = (1-Pe) * N(X,Y) / N(X), \quad if\ N(X,Y) > 0 \quad (4)$$
$$= Pe * (N(Y)+1) / (Nt+10000), \quad if\ N(X,Y) = 0$$

$$Pe = (Ns(X)+1) / (N(X)+2) \quad (5)$$

其中 N(X,Y) 表示在訓練語料中 Y 緊接著 X 出現的次數，N(Z)表示 Z 在訓練語料中出現的次數，Ns(X)表示具有 N(X,Y)=1 之不同的 Y 的個數，Nt表示訓練語料的總字數，而Pe則表示逃脫機率。另外，關於 P(X) 之數值，本研究是以 X 之詞頻除以一個常數值 1,000,000 來估計。

四、學習部門

由於系統裡原有之詞典不可能將各行各業會用到的詞語都收錄進來，並且使用者才最清礎那一個字、詞會經常出現於要輸入的文章中，因此，本部門提供了線上詞彙學習功能，以讓使用者將新的詞彙插入系統之詞典，也可作消除舊詞的處理，此外，不管新詞或舊詞，也都可學習其詞頻和在同音詞中的排名次序，與一般以注音輸入中文的系統，只能學習新詞(不含詞頻)的情況有很大的不同。如果能夠修改一個詞的詞頻，則使用者可以控制音轉字部門之動態規劃處理的選字動作，例如可把某一不想要的同音詞之詞頻降低很多，而讓本次要輸入的文章的一個常用到的同音詞之詞頻增加，如此，就可減少音轉字後的錯誤字。

此外，本部門也提供了近形字群線上建立的功能，當建立近形字群後，就可使用近形字查詢的功能。近形字的觀念是要來解決一些不常用的字，因不會唸而無法輸入的問題，也就是說，如果我們將一些部首不同而主體部分相同的字(如：者、著、賭、堵、都、諸、睹… 等)建立出一個近形字群的關係放在詞典裡，則當要輸入"睹"但不會唸時，可先輸入"者"或同一群裡的其它近形字，再使用近形字查詢功能，將"者"或其它近形字換成"睹"，這種操作程序類似於同音字查詢的操作，用於更改自動音轉字時被轉換錯誤的字。近形字之觀念，我們未曾在其它以注音輸入中文的系統中看過，因此，這樣的觀念算是本系統一項重要特色，它對採用注音之中文輸入方法的推廣有很大的幫助，因爲許多潛在的使用者(如中小學生)當他不知道所要輸入字的注音時，就可先輸入所要輸入字的形狀相似字，然後按近形字查詢的按鍵，就可由此字的近形字群中，去選取他所要輸入的中文字。

五、編輯與輔助部門

本部門提供了一些文字編輯功能及其它的輔助功能，以便讓使用者在輸入一段或一篇文章後(或任何時候)，都可暫停輸入，而去對先前所輸入之文字作編輯或有關的處理動作，先前輸入的文句顯示於螢幕上的文字區，而緩衝區則用於顯示目前正在輸入的句子，使用者也可在緩衝區操作部份的編輯與輔助功能。本部門所提供的編輯功能包含鍵盤上之編輯按鍵所代表的之外，還有區塊(block)之設定、複製、與消除，以及讀取、寫出文字資料檔(可選擇檔案裡是否要存全篇字的注音資料)等。由於本部門提供的存檔、讀檔功能，可把文字

對應的注音資料也一起寫出及讀入，所以可等下一次執行本系統時再來操作同音字查詢功能，去改正轉換錯的字，此時當然也可由原輸入者以外的人來操作。

除了一般的編輯功能之外，本部門也提供了一些很有用的輔助功能，亦即它能夠在收到特定的按鍵後去呼叫：同音字、詞查詢之功能(依據系統內保存的注音資料，查無注音資料時會要求使用者輸入)；近形字查詢的功能；自動查詞典以設定全篇中文字之注音的功能(即自動字轉音，可用以設定由其它軟體輸入的文字的注音)；全篇字的注音資料轉換成Big-5碼並存到檔案去；全篇自動音轉字的功能；以及其它的功能。所以，本部門已將文句編輯功能與中文輸入有關之輔助功能整合在一起，使得使用者可隨意變換去螢幕上的緩衝區進行輸入，或者去文字區編輯文句與執行輔助功能，這種操作上的方便性，一般的中文輸入系統並沒有提供。

六、音轉字實驗

當應用馬可夫模型來解決自動音轉字的問題時，在實做的考慮下，可被選用的具體模型包括MW0(以詞爲處理單位，但不作預測)、MW1(以前一個詞來預測下一詞)、MC1(以前一個字來預測下一字)、或MC2(以前二個字來預測下一字)等。可是，再考慮本系統提供的線上新詞學習功能時，MW1、MC1、與MC2模型的應用就碰到困難了，因此，我們才提出一種稱爲 MW0/MC1 之複合模型，來把原先的MW0與MC1模型加以結合運用。

這一節我們就拿 MW0、MC1、及MW0/MC1等模型來進行音轉字之實驗，看所提出模型的音轉字正確率，和基本模型MW0與MC1的正確率有無差別，爲了使情況單純，在實驗進行中並不做線上新詞加入的處理。實驗所用的測試文章，一篇是取自晚報社論，共694個音節，另外一篇則是取自國語月刊的兒童故事，共501音節，這兩篇文章未被用於訓練音轉字模型。系統裡的詞典約有 45,000 個詞，它們的詞頻可用以估計第三節各式子裡的機率項 P(X)，至於條件機率項 P(Y|X) 的估計，我們使用了大約 230 Kbytes 的國小國語課本之課文，及大約 170 Kbytes 的報紙社論文章來統計出那些有關的計次參數的數值，以便代入式子(4)及(5)去估計所要的條件機率值。另外，關於複合模型的機率值計算，經過初步實驗之觀察後，我們令式子(1)裡的加權常數 w 的數值爲 2。

將測試文章所對應的音節注音送給各個模型去作自動音轉字處理，然後比對轉換結果與原始文章，我們就得到了如表1所示的音轉字正確率值。由此表可知，各種模型的轉換率都可達到92.6以上，其中複合模型(MW0/MC1)對兩篇測試文章各得到93.6與94.4之轉換率，都比MW0模型的92.6與93.8來得高，這說明複合模型可以用來改進MW0模型；另外，複合模型和MC1模型比起來則互有

表1 數種模型之音轉字正確率

音轉字方法 / 測試文章	MW0	MC1	MW0/MC1
國語月刊故事	92.6	94.8	93.6
晚報社論	93.8	92.8	94.4

高下，但如果以兩篇文章中轉換率較低者來比時，則複合模型的93.6要比MC1的92.8要高一些，這說明複合模型的轉換率較爲穩定。不過，由於所用的測試音節數量並不夠多，所以前述的論點還需進一步的實驗來驗證。

七、結語

我們在設計、製作本中文輸入系統時，是採取實用的觀點，希望所製作的系統可以眞槍實彈給使用者去操作使用，所以花了許多時間精力於寫作與偵錯程式，以提供大大小小之各種功能。雖然我們的系統必需進入一個基礎中文系統(如倚天中文系統)後才可操作使用，但是我們的系統和基礎中文系統是相輔相成而不會衝突的，即可用我們的系統來輸入中文外，也可切換使用基礎中文系統所提供的中文輸入方法。

本系統採用之宜韻注音鍵盤，考慮了三項鍵盤設計的重要準則，即鍵盤效率，人體工學原則，及符號至按鍵對應的規律性，其中，鍵盤效率所指的是輸入一個音節的平均按鍵次數；而人體工學原則是要儘量減少手指頭的運動量，以避免疲勞；至於符鍵對應規律性的目標是，讓使用者很輕鬆地在建立符號和按鍵位置的聯想對應關係。

關於自動音轉字的處理，本系統採用了新提出的複合馬可夫語言模型的作法，這樣的模型除了可支援線上新詞學習的功能外，也兼顧了句子裡相連兩詞間的相關性。初步測試實驗的結果顯示，複合馬可夫模型可改進原來 MW0 模型之音轉字正確率，並且比原來的 MC1 模型穩定，不過這還需進一步的實驗來驗證。如果使用者能夠適當地操作線上新詞登錄功能，則整體的音轉字正確率應該會再提高。

本系統也提供了近形字群線上建立的功能，以及近形字查詢的功能，這是考慮許多潛在的使用者(如中小學生)可能不知道所要輸入字的注音，這時他就可先輸入所要輸入字的形狀相似字，然後透過近形字查詢的功能，去選取他所要輸入的中文字，所以，近形字的觀念非常有助於以注音輸入中文之系統的推廣。

參考文獻

[1] Microsoft Corporation, Microsoft Windows 中文版，1992。

[2] 倚天資訊有限公司，倚天神雕筆手寫辨識系統使用手冊，1992年。

[3] 蒙恬科技有限公司，蒙恬中國筆使用手冊，1991年10月。

[4] 范欽雄、李豐壽，「應用類神經網路辨認常用中文字5401字」，全國計算機會議論文集(嘉義)，第619-627頁，1993。

[5] 黃雅軒等，「印刷體光學中文字形辨識系統」，電子發展月刊，第141期，第16-26頁，民國78年9月。

[6] Lee, Lin-shan, chiu-yu Tseng, Hung-yan Gu, *et al.*, "Golden Mandarin(I) -- A Real-time Mandarin Speech Dictation Machine for Chinese Language with Very Large Vocabulary", IEEE Trans. Speech and Audio Processing, pp. 158-179, 1993.

[7] 倚天資訊有限公司，倚天中文系統使用手冊，1992年5月。

[8] 松下電器開發有限公司，漢音詞彙輸入法使用手冊，1991年12月。

[9] 長諾資訊圖書股份有限公司，國音輸入法，1993年5月。

[10] 古鴻炎，「一個同時考慮鍵盤效率人體工學原則及符鍵對應規律性之國語注音輸入鍵盤的設計」，電工雙月刊，第35卷，第2期，第123-132頁，1992年4月。

[11] Gu, H. Y., A Study on a few Relevant Problems about Machine Dictation of Mandarin Speech, Ph. D. Dissertation, Department of CSIE, National Taiwan University, Jan. 1990.

[12] Gu, H. Y., C. Y. Tseng and L. S. Lee, "Markov Modeling of Mandarin Chinese for Decoding the Phonetic Sequence into Chinese Characters", Computer Speech and Language, Vol. 5, No. 4, pp. 363-377, 1991.

[13] Kuo, J. J., J. H. Jou, M. S. Hsieh, and F. Maehara, "The Development of New Chinese Input Method -- Chinese Word-string Input System", Proceedings of International Computer Symposium (Tainan, Taiwan), pp. 1470-1479, 1986.

[14] Lin, M. Y. and W. H. Tsai, "Removing the Ambiguity of Phonetic Chinese Input by the Relaxation Technique", Computer Processing of Chinese and Oriental Lanuages, pp. 1-24, 1987.

[15] 季震寰，結合詞與統計的注音中文輸入系統，國立台灣大學資訊工程系碩士論文，80年7月。

[16] Hsieh, M. L., T. T. Lo and C. H. Lin, "Grammactical Approach to Converting Phonetic Symbols into Characters"， Proceedings of National Computer Symposium (Taipei)，pp. 453-461，1989.

[17] Ross, S. M., Introduction to Probability Models, third edition, Academic Press, Inc., 1985.

[18] Katz, S. M., "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer", IEEE trans. Acoust., Speech, and Signal Processing, pp. 400-401, March 1987.

[19] Witten, I. H. and T. C. Bell, "The Zero-Frequency Problem: Eetimating the Probabilities of Novel Events in Adaptive Text Compression", IEEE trans. Information Theory, Vol. 37, pp. 1085-1094, 1991.

Quantitative Corpus Analyses of Character Errors in Primary School Students' Chinese Writings in Taiwan

Yuangshan Chuang, Ph.D.
Department of English
National Kaohsiung Normal University
T1730@nknucc.nknu.edu.tw

ABSTRACT

This paper quantitatively describes differences of character error patterns in terms of the psychological effects of form (F), sound (S), and meaning (M) on primary school students' substituting wrong characters for right ones. The corpora from which the character errors were collected consisted of 2,104 and 1,998 compositions written respectively by 357 second, third, and fourth grade and 301 fifth and sixth grade students in the first semester of the 1993 academic year. The error patterns were partitioned into 7 categories: F, S, M, FS, FM, SM, and FSM, the first three of which formed one significance test group and the second three of which composed another. The one-factor repeated measures ANOVA model was used in the significance tests. The results of this study showed significant differences at the level of .01 in the effects between sound and form and also between sound and meaning. There were also significant discrepancies between form-sound and form-meaning, between form-sound and sound-meaning, and between sound-meaning and form-meaning as well. When we separated the character errors into the second, third, and fourth grade corpus and the fifth and sixth grade corpus and conducted a significance test for each, both of the tests rendered the same results as that for the entire corpus composed of the two groups. This indicated that the grade factor did not significantly contribute to the differences in the effects of character features.

INTRODUCTION

In Taiwan, character errors are prevalent in primary school students' Chinese writings. It is of significance to analyze character errors so that findings about them can be applied to avoid students writing characters incorrectly. Although researchers such as Chen (1978) and Lin (1983) usually divide error types on the basis of form (F), sound (S), and meaning (M), not much research has been performed to classify error patterns statistically as well as psychologically. Moreover, until recently researchers such as Tseng and Hoosain have laid their emphases on language recognition and comprehension, neglecting language production (陳, 1993, p. 27). Therefore, it is of great importance to collect students' writings and analyze character errors in them quantitatively.

We randomly selected 2,104 and 1,998 pieces of composition written respectively by 357 second, third, and fourth grade students and 301 fifth and

sixth students in the first semester of the 1993 academic year. The items under examination were the characters in the students' compositions written in the first semester of the 1993 academic year.

According to Tang (1989, p. 20), "character" is an orthographic unit that can take up a square in a piece of draft paper and may or may not have meaning. For example, 樹 "shu4" is meaningful while 徘 "pai2" has to be used together with 徊 "huai2" to form a complete meaning. "Morph" is a semantic unit that must have meaning but may or may not be used independently, and "word" is a syntactic unit that functions meaningfully and may contain only one or more than one character. In this study, tabulation of character errors was conducted based on the unit of character.

Significance of the Study

Character errors may unpleasantly hinder written communication. For example, Emperor Kang Si wrongly substituted 近年 "jin4 nian2" (recent years) for 今年 "jin1 nian2" (this year) (劉, 1992, p. 29), which was sure to greatly distort his actual meaning conveyed to his subjects. They may also prevent students from adequately acquiring Chinese, since characters composing words play an essential part in the acquisition of the four language skills: listening, speaking, reading and writing.

Research Questions and Hypotheses

This study will answer the following two questions:

1. Are the effects of Chinese character features different in influencing students in making character errors in their writings?
2. Does the grade factor contribute to differences in psychological effects of Chinese features?

These questions are raised on the basis of the following hypotheses:

1. There will be significant differences in the psychological effects of the Chinese character features.
2. The grade factor will significantly influence the psychological effects.

Research Design

To test the hypotheses postulated above, frequencies of character errors based on character features were tabulated respectively for the STFG and FSG students and for the entire corpus. The inappropriately substituted characters were partitioned into seven categories: form (F), sound (S), meaning (M), form-sound (FS), form-meaning (FM), sound-meaning (SM), and form-sound-meaning (FSM). The categorization was based on the relation between the substituted character and the replaced counterpart in terms of character features

and also the context in which the character appeared. For instance, since the character 以 in 以經 was produced with sound interference, it was therefore listed under the category of S. What is more, the character 輝 in 輝棒落空 belonged to the category of FS because 輝 and 揮 were related in both form and sound.

After the partitioning of the errors and calculation of them in each category were completed, the one-factor repeated measures ANOVA model was applied to examine the discrepancy between effects of character features on character errors for the one-feature group of F, S, and M and the two-feature one of FS, FM, and SM. Since there was a significant difference among the means both in the entire corpus and in each of the separate corpora, six Scheffe post hoc multiple comparisons of means were conducted to examine which pair or pairs of means were significantly different from each other.

RESEARCH RESULTS AND DISCUSSION

Character Error Distributions

The character error frequencies were computed respectively for the entire corpus and for each of the separate corpora. Table 1 presents frequency distributions derived from calculation of the inappropriately substituted characters based on character features. The figures indicate that the feature of sound in the single-feature group and that of form-sound in the double-feature group might exert the highest interference with students' character production in their writings. However, significance tests had to be conducted to see if the effects of character features were significantly different in interfering with students' character production.

Table 1
Frequency Distributions and Percentages for the Character Errors in the Entire Corpus

Features	STFG	Percentage	FSG	Percentage	Total	Percentage
F	238	8.756	216	10.390	454	9.464
S	1,653	60.817	804	38.672	2,457	51.220
M	126	4.636	153	7.360	279	5.816
FS	531	19.536	616	29.630	1,147	23.911
FM	13	0.478	7	0.337	20	0.417
SM	105	3.863	162	7.792	267	5.566
FSM	52	1.913	121	5.820	173	3.606
Total	2,718	100.00*	2,079	100.00*	4,797	100.000*

*The total percentage was rounded.

Errors Existing in Both Groups of Students' Writings

There were 1,166 different pair patterns of character errors in the STFG school students' compositions, of which 322 pairs occurred in the FSG students'. That is to say, 27.616 (322/1,166) percent of the same patterns of errors found in the STFG students' compositions were made again by the FSG students. If we regard each character error as individual, there were 2,718 errors by the STFG students and 2,079 errors by the FSG students. Among the 2,718 errors, 340 appeared two times or more, in which 171 (50.294 percent) reappeared in the FSG students' writings; 178 occurred three times or more, in which 117 (65.730 percent) reoccurred; and 111 took place four times or more, in which 83 (74.775 percent) took place again.

Inferential Statistics

Since the numbers of compositions collected from each class differed, we converted the raw scores into adjusted scores by dividing each error score by the total error score of its character feature group in that class and multiplying the decimal by 100. For example, the raw score 26 in class CS 2-8 would become 10.788 (26/241x100) after it was converted since the total error score of the single-feature group was 241. Then the one-factor repeated measures ANOVA model was applied to test the first hypothesis, using the adjusted scores. When a significant difference was found, the Scheffe post hoc procedure was conducted to examine which pair or pairs of means were significantly different. And the second hypothesis was tested by comparing the significance test results for the STFG and FSG corpora.

In this section, inferential statistics will be presented. Character features were used as predictor of error scores in the one-factor repeated measures ANOVA model and the multiple comparison procedures (Lomax, 1992, pp. 221-232 and pp. 143-144; 林, 1988, pp. 283-287). These significance tests were performed to find whether different types of character features contributed to significantly different interference with students' character errors in their writings.

Results of ANOVAs and the Scheffe Procedure

Tables 2 through 5 illustrate the results of the ANOVAs and the Scheffe procedure for the tests of significant differences among the features of form, sound, and meaning and those of form-sound, form-meaning, and sound-meaning. There were significant results for both the single-feature group and the double-feature group. In order to find out which pair or pairs of means contributed significantly to the variation, the Scheffe procedure was implemented. From Table 3, we can find that there were significant differences

Table 2
ANOVA for the Single-Feature Group in the Entire Corpus

Source	df	Sum of Squares	Mean Squares	F-test	p value
Between classes	19				
Within classes	40	54809.795			
treatments	2	52353.676	26176.838	404.997*	.0001
residuals	38	2456.119	64.635		
Total	59	54809.795			

*p < .01

Table 3
Scheffe Procedure for Mean Differences Based on Single Features in the Entire Corpus

	$\bar{Y}.1$ (meaning)**	$\bar{Y}.2$ (form)**	$\bar{Y}.3$ (sound)**
$\bar{Y}.1 = 9.415$	--	6.216	65.538*
$\bar{Y}.2 = 15.631$		--	59.322*
$\bar{Y}.3 = 74.954$			--

*p < .01 **$\bar{Y}.1$, $\bar{Y}.2$, and $\bar{Y}.3$ stand for the means.

between sound and form and between sound and meaning, but not between form and meaning, with the feature of sound offering the highest effect. Table 5 shows that there were significant differences between all the three pairs of means for the double-feature group, with the feature of form-sound exerting the biggest influence.

Table 4
ANOVA for the Double-Feature Group in the Entire Corpus

Source	df	Sum of Squares	Mean Squares	F-test	p value
Between classes	19				
Within classes	40	70395.914			
treatments	2	67558.885	33779.442	452.452*	.0001
residuals	38	2837.029	74.659		
Total	59	70395.914			

*p < .01

Table 5
Scheffe Procedure for Mean Differences Based on Double Features in the Entire Corpus

	$\bar{Y}.1$ (FM)	$\bar{Y}.2$ (SM)	$\bar{Y}.3$ (FS)
$\bar{Y}.1 = 1.500$	--	17.271*	78.229*
$\bar{Y}.2 = 18.771$		--	60.957*
$\bar{Y}.3 = 79.729$			--

*$p < .01$

Discrepancies in Effects of Character Features in Terms of Grade Differences

In order to see whether the grade factor contributed significantly to the discrepancies in effects of character features on character errors, we separated the entire corpus into two sets of data. One of them consisted of the character errors made by the STFG students and the other was composed of those found in the FSG students' compositions. The one-factor repeated measures ANOVA model and the Scheffe procedure were conducted again for each of the two grade levels. Tables 6 through 13 indicate the same results for both the STFG corpus and the FSG data as those for the entire corpus. That is to say, it was not the grade factor that caused character features to significantly function differently in interfering with students' character errors.

Table 6
ANOVA for the Single-Feature Group in the STFG Corpus

Source	df	Sum of Squares	Mean Squares	F-test	p value
Between classes	9				
Within classes	20	35539.018			
treatments	2	35203.995	17601.997	945.713*	.0001
residuals	18	335.023	18.612		
Total	29	35539.018			

*$p < .01$

Table 7
Scheffe Procedure for Mean Differences Based on Single Features in the STFG Corpus

	$\bar{Y}.1$ (meaning)	$\bar{Y}.2$ (form)	$\bar{Y}.3$ (sound)
$\bar{Y}.1 = 6.452$	--	5.418	75.225*
$\bar{Y}.2 = 11.870$		--	69.807*
$\bar{Y}.3 = 81.677$			--

*p < .01

Table 8
ANOVA for the Double-Feature Group in the STFG Corpus

Source	df	Sum of Squares	Mean Squares	F-test	p value
Between classes	9				
Within classes	20	35847.397			
treatments	2	34622.052	17311.026	254.295*	.0001
residuals	18	1225.345	68.075		
Total	29	35847.397			

*p < .01

Table 9
Scheffe Procedure for Mean Differences Based on Double Features in the STFG Corpus

	$\bar{Y}.1$ (FM)	$\bar{Y}.2$ (SM)	$\bar{Y}.3$ (FS)
$\bar{Y}.1 = 1.844$	--	15.810*	78.657*
$\bar{Y}.2 = 17.654$		--	62.847*
$\bar{Y}.3 = 80.501$			--

*p < .01

Table 10
ANOVA for the Single-Feature Group in the FSG Corpus

Source	df	Sum of Squares	Mean Squares	F-test	p value
Between classes	9				
Within classes	20	19270.777			
treatments	2	18512.348	9256.174	219.679*	.0001
residuals	18	758.429	42.135		
Total	29	19270.777			

*p < .01

Table 11
Scheffe Procedure for Mean Differences Based on Single Features in the FSG Corpus

	$\bar{Y}.1$ (form)	$\bar{Y}.2$ (meaning)	$\bar{Y}.3$ (sound)
$\bar{Y}.1 = 12.378$	--	7.014	55.851*
$\bar{Y}.2 = 19.392$		--	48.838*
$\bar{Y}.3 = 68.230$			--

*p < .01

Table 12
ANOVA for the Double-Feature Group in the FSG Corpus

Source	df	Sum of Squares	Mean Squares	F-test	p value
Between classes	9				
Within classes	20	34548.517			
treatments	2	32976.107	16488.053	188.745*	.0001
residuals	18	1572.411	87.356		
Total	29	34548.517			

*p < .01

Table 13
Scheffe Procedure for Mean Differences Based on Double Features in the FSG Corpus

	$\bar{Y}.1$ (FM)	$\bar{Y}.2$ (SM)	$\bar{Y}.3$ (FS)
$\bar{Y}.1 = 1.155$	--	18.733*	77.801*
$\bar{Y}.2 = 19.889$		--	59.067*
$\bar{Y}.3 = 78.956$			--

*$p < .01$

CONCLUSION

Chinese character features were analyzed in this study. There were three sets of character error data collected for significance tests to see how character features functioned in influencing the students in making errors in Chinese characters. The tests were performed for the two hypotheses posited at the beginning of this study. Not both hypotheses were justified by the significance test results.

First, the results of this study found significant differences in the effects between sound and form and between sound and meaning. There were also significant differences between form-sound and form-meaning, between form-sound and sound-meaning, and also between sound-meaning and form-meaning.

The second hypothesis asked for an examination of differences in effects of character features in terms of grade differences. The significance tests respectively for the STFG and FSG corpora were implemented. From the viewpoint of significance, the results for character features were consistent for each of the two grade levels and the two levels combined. That is to say, the grade factor did not affect the students' character errors significantly. Therefore, the hypothesis was rejected.

It is expected that this study will contribute to the unraveling of mystery behind Chinese character errors and offer useful information for language researchers, curriculum developers and textbook editors, with which they can better understand the nature of Chinese character errors and therefore prepare more effective Chinese teaching materials and methods for students who learn Chinese as their mother tongue. It is also hoped that the findings may contribute to the design of more effective learning materials for learners of Chinese as a foreign or second language and in turn to the improvement of the students' Chinese acquisition.

REFERENCES

Lomax, R. G. 1992. Statistical concepts. White Plains, N.Y.: Longman.

林清山 (Lin). (1988). 心理教育統計學. 臺北: 東華.

林瑞端 (Lin). (1983). 兒童錯別字研究. 台灣省國民中小學專題研究報告第三輯. 台灣: 台灣省政府教育廳.

湯廷池 (Tang). 1989. 漢語旳〈字〉、〈詞〉、〈語〉、〈語素〉. 華文世界, 53, 18-29.

陳光政 (Chen). 1978. 常用詞彙國音誤讀資料研究報告(一)--如何校訂別字. 台灣: 高師院國文系.

陳振宇 (Chen). 1993. 一些國語的自然語誤及其分類. 華文世界, 69, 26-41.

劉家駒 (Liu). (1992). 康熙皇帝的啟蒙教育--由其硃批中的錯別字談起. 故宮文物月刊, 109, 26-37.

CHINESE-WORD SEGMENTATION BASED ON MAXIMAL-MATCHING AND BIGRAM TECHNIQUES

Luk Wing Pong, Robert (陸永邦)
Department of Chinese, Translation and Linguistics
City Polytechnic of Hong Kong
Email: CTRWPL92@CPHKVX.CPHK.HK

ABSTRACT

One of the most simple and accurate Chinese-word segmentation technique is maximal-matching. However, its performance depends on the coverage of the list of words which are usually derived from a general dictionary. When it is directly applied to segment technical articles instead of general news articles, the error rate degraded significantly from 1.2% (as in the literature) to 15%. This is an important problem in two respect. First, usually the domain-specific terms are not readily available on computer. These terms have to be entered manually by expert or they can be detected automatically from thematic corpora. Second, if corpus analysis is applied to supplement information for the design and development of text processing systems, these analysis depend on the correct word segmentation of these corpora of technical articles. In this paper, we propose to combine the maximal-matching and bigram techniques in Chinese-word segmentation for detecting words in thematic corpora, where both techniques overcome each other's short coming. The Hong Kong Basic Law was selected as a representative technical article for evaluation because it has a fair amount of technical terms, compound nouns and names. The segmentation performances of the maximal-matching, bigram and the combined techniques are compared. The combined technique was able to achieve 33% improvement in segmentation performance and identify 33% of the terms in the Basic Law.

I. INTRODUCTION

Thematic corpora are compiled to enlarge the scale of the sub-language approach [1] in text processing systems (e.g. machine translation and text retrieval), on the one hand to supplement empirical information for the design of these systems and on the other hand to evaluate these systems with authentic data. However, techniques for corpus analysis may not be appropriate for thematic corpora because they were developed to analyze general articles which are sampled across various domains. By contrast, thematic corpora sample articles of a specific domain and these articles tend to be technical, for example, constructing a machine translation system for financial reports or text retrieval system for constitutional law. An important case in point is Chinese-word segmentation which is an elementary stage of any Chinese text processing systems. For general corpora, it has been acknowledged that names and proper nouns constitute major errors in word segmentation [2,3] even though the amount of segmentation error is small, typically between 1% and 2%. However, the amount of error may increase significantly with thematic corpora since technical terms and compound nouns are more likely to occur for technical articles. Although domain-specific dictionaries (e.g. dictionary on computers) are available, the representativeness of these entries have to be evaluated using the thematic corpora. These entries are usually entered manually because current OCR technologies are not as cost-effective as professional typists, given that errors occur frequently when character size changes as in many dictionaries. An alternative is to extract a tentative list of technical terms or proper nouns from thematic corpora and the list is verified using a dictionary or extended manually using a concordance program [4].

Extracting the list is similar to detecting proper nouns and technical terms as in text retrieval for English where strongly associated words are grouped together, depending on their co-occurring frequencies or mutual information within a specified context [5]. Syntactic patterns are also used to eliminate improbable cases [6]. Detection for Chinese is simpler than for English since Chinese terms and proper nouns tend to be a sequence of consecutive characters. Detection of two-character words have already been reported in [7,8] but technical terms and proper nouns usually have more than 2 characters, particularly those that are translated. We are unaware of any report in the literature about the effectiveness of detecting two-character words in improving the segmentation performance.

The aim of this paper is to address the problem of word detection for improving the performance of Chinese word segmentation of thematic corpora. We combined both maximal-matching [9] and bigram [7]

techniques because they complement each other and they are relatively inexpensive and simple to implement compared with the relaxation [10], adaptive statistical [11] and competitive neural network [12] techniques. In addition, the combined technique does not need to estimate positions of error occurrence for each thematic corpus as for the adaptive statistical technique.

The basic idea of the combined technique is to use a list of words to match with the input clause from left-to-right. A new segmentation and matching position is found at the end of the longest matched word. Since segmentation errors are due to the coverage of the list of words, typically the terms tend to be over-segmented into smaller ones. For example, the term, 中華人民共和國 (People's Republic of China), is over segmented in the following clause: /根據/中/華/人民/共/和/國/憲法/第/三十一/條/的/規定/，/.

To reduce the amount of over-segmentation, adjacent single-character words are grouped using the bigram technique after maximal-matching. These characters are combined if their mutual information (MI) or co-occurrence frequencies (CF) are greater than a threshold. MI and CF are estimated from the thematic rather than a general corpus because the estimated values should be biased to the thematic corpus. The CF is estimated as the frequency of occurrence of character A immediately before B (i.e. f(A,B)) whereas MI is estimated as log (p(A,B) / (p(A) * p(B)) where p(A,B) (= f(A,B)/N where N is the sum of all CF's) is the estimated probability of character A occurring immediately before B, p(A) and p(B) are the estimated probabilities of character A and B in the corpus, respectively. The threshold is defined by the top N% of the MI or CF distributions. Typically, the MI distribution appears symmetrical and unimodal (figure 1) but the CF distribution decreases with increasing CF and log CF. Although the bigram technique can be modified to identify words of arbitrary length, many non-words are detected. By combining with maximal-matching, the bigram technique can only operate in certain parts of the thematic corpus, reducing the number of non-words detected.

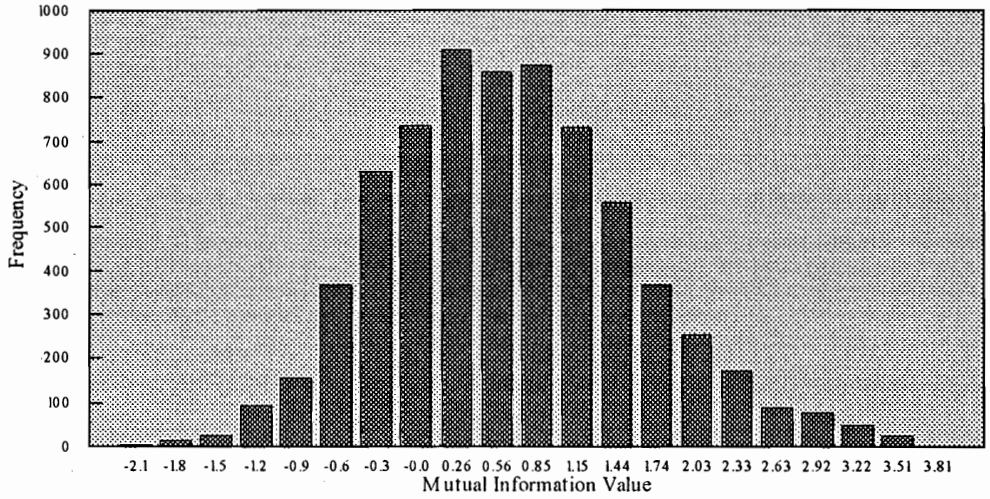


Figure 1: Frequency distribution of MI estimated from the Hong Kong Basic Law. Note that the distribution appears like a normal curve as found in large corpora and the mean position is slightly larger than zero (i.e. two characters are independent).

In the rest of this paper, we discuss how the segmentation programs are evaluated. Next, we compare the performance of maximal-matching between words extracted from a general corpus and those extracted from the manually-segmented text. We show that the bigram technique is equivalent to the nearest-neighbor (NN) clustering. We report the effect of adjusting the threshold (or percentage quartile) on its segmentation and word-identification performances. Finally, we report the result of the combined technique.

II. EVALUATION

We chose the Hong Kong Basic Law [13] as our test data because it

(a) contains a fair amount of proper names, technical terms and compound nouns (see figure 2);
(b) is a typical technical text (in law) which is suitable for machine translation and corpus analysis for humanities and law research;
(c) is large enough for estimating the mutual information in data-exploration since the

distribution of mutual information appear like a normal curve as found in large corpora (figure 1);

(d) is manually segmented and accessible on the computer [14].

第十一條
根據中華人民共和國憲法第三十一條，
香港特別行政區的制度和政策，
包括社會、經濟制度，
有關保障居民的基本權利和自由的制度，
行政管理、立法和司法方面的制度，
以及有關政策，
均以本法的規定爲依據。
香港特別行政區立法機關制定的任何法律，
均不得同本法相抵觸。

Figure 2: An extract of the Hong Kong Basic Law. Terms can already be found, such as the Hong Kong Special Administrative Zone, People's Republic of China and Economic System.

The Basic Law (Figure 2) is different from a general corpus, such as the PH corpus [4] of general news articles. Only 1059 different words appeared in both the Basic Law (2,028 different words) and the PH corpus (42,613 different words), representing 52% and 2.5% overlap, respectively. Figure 2 shows the variation of percentages of word overlap in the Basic Law with the length of the word (i.e. the number of characters). Single-character and two-character words have relatively high percentages of overlap compared with longer words because technical terms and compound nouns tend to be long (> 2 characters), particularly with terms that originated or translated from foreign languages like English, (e.g. aspirin as 阿司匹靈 or tort as 民事過失).

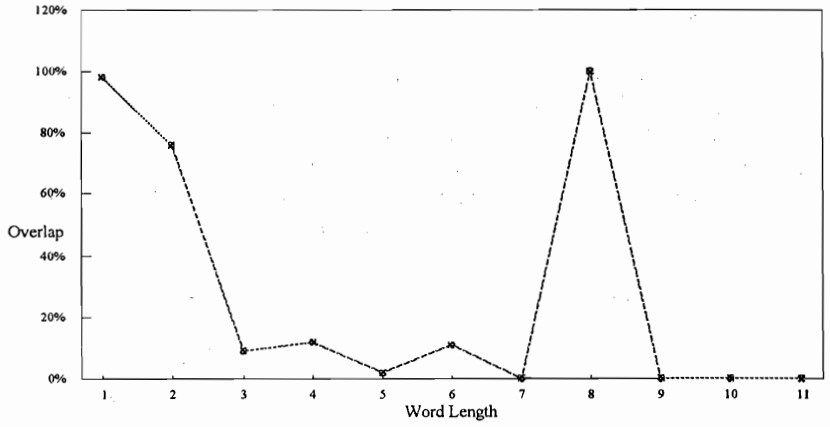


Figure 2: Percentages of words that occur in both the Hong Kong Basic Law and the PH corpus. Note that the percentages of the 8-character words are not accurate because there is only a single 8-character word in the Hong Kong Basic Law.

The Basic Law was segmented by a graduate student in applied linguistics. She was told to segment words as long as possible. For example, the compound noun, 香港特別行政區 (Hong Kong Special Administrative Zone), is considered as a single word rather than its constituents, /香港/特別/行政/區/ (/Hong Kong/ Special/ Administrative/ Zone/). Segmentation markers are placed before and after 的 (de) whenever it is used as an adjective suffix, for example /中國/的/領土/ (/China/ de/ Territories/), since 的 is very productive, creating many new words if it is combined with its adjective root. Unlike [15], numbers are grouped together as a unit, for example /一九八四/年/十二/月/十九/日/ (/1984/Year/December/Month/Nineteenth/Day/).

The segmentation performance is measured by comparing the automatically and manually segmented clauses of the same text. Given that clause i has N_i characters, the maximum amount of segmentation errors is N_{i-1}. Thus, the normalized segmentation error for every clause is $NE_i = E_i/N_i$ where E_i is the amount of segmentation errors. The mean segmentation error is defined as follows where there are k clauses: $E = (\Sigma_i NE_i) / k$.

A single state transducer is used to determine E_i. Initially, E_i is set to zero and the transducer begins at the left-most position on both clauses. The transducer has only two types of actions depending on whether the input symbols are segmentation markers or not. If one clause has segmentation marker but the other does not, then the transducer increments E_i by one and it advances beyond the position of the segmentation marker but remains at the same position on the other clause. If the transducer encounters either segmentation markers or none on both clauses, then there are no segmentation errors and it moves to the following positions of both clauses.

The amount of over-segmentation can be measured simply as the number of automatic segmentation markers, $N_{a,i}$, minus the number of manual segmentation markers, $N_{m,i}$, in clause i, excluding the markers at the beginning and the end of the clause. The over-segmentation is normalized to $N_{m,i}$ because that is the desired number of segmentation. The mean over-segmentation is defined as: $O = [\Sigma_i (N_{a,i}-N_{m,i})/N_{m,i}] / k$.

If O is positive, then there are over-segmentations and vice versa. If O is zero, then the amount of manual segmentation is approximately the same as the automatic. However, we need to examine E to determine whether these segmentations are accurate.

Apart from segmentation, the bigram technique identifies new words. The identification performance (C/B) can be measured as the percentage of identified words, $W_{b,i}$, in the list of different words, W_b, extracted from the Basic Law. However, this measure does not indicate whether the bigram is detecting words that are not in the Basic Law. Thus, another percentage (C/T) defines the ratio between the identified words that are in the Basic Law and the number of identified words.

III. MAXIMAL-MATCHING TECHNIQUE

Maximal-matching depends on the coverage of the dictionary. We compared the list of words extracted from the Basic Law and its subset that also occurred in the PH corpus. The first list achieve a low segmentation error of 1.2% compared with 15% using the second list. Consequently, the clause accuracy for the first list (82%) is much better than the second (28%). This is quite surprising since the words extracted from the PH corpus are derived from a general Chinese dictionary [16] of about 56,000 words. The first list under-segments the Basic Law (i.e. 3%) where as the other over-segments (6%).

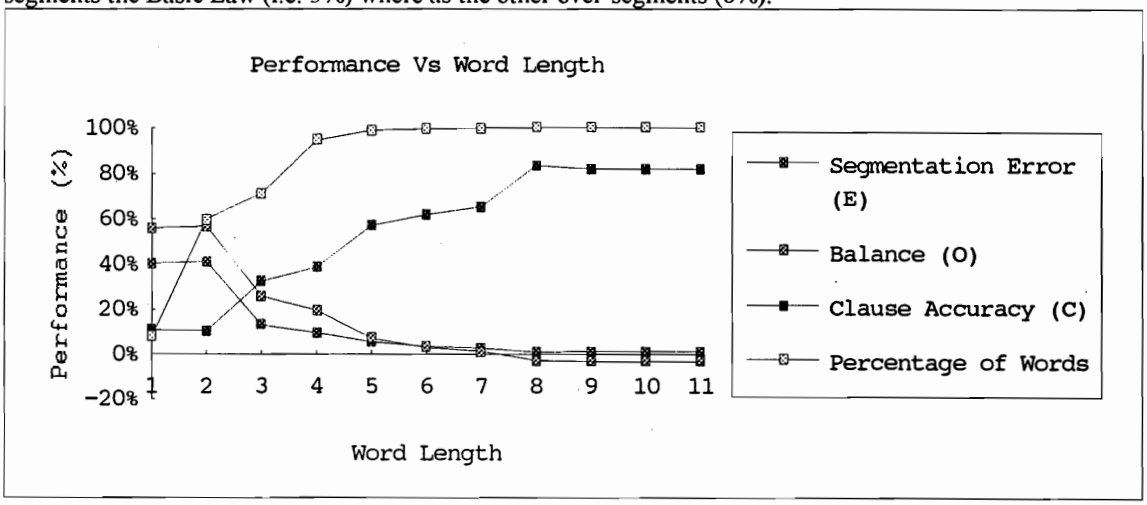


Figure 3: Variation of segmentation performance with the different words extracted from the Hong Kong Basic Law, up to certain lengths. Key: E represent the segmentation error, O measures the over-segmentations, C is the clause accuracy and W is the amount of words up to a particular length.

Since the maximal-matching uses the longest matched word, we were interested in the variation of segmentation performance with the list of words up to a particular length. The segmentation error reduces as longer words are included (Figure 3). However, the addition of 2-character words do not in general improve the segmentation performances of the single-character words. Longer words have to be identified for

significant improvement. The amount of over-segmentation reduces near to zero when words of length up to 7 characters are used in segmentation. Longer words will moderately make maximal-matching under-segments but improving the clause accuracy significantly.

IV. BIGRAM TECHNIQUE

Previous work [7,8] grouped two adjacent characters as a two-character word. However, we showed that detecting only two-character words hardly improve segmentation performance in the last section. Thus, we extended the idea to detect words of arbitrary length by grouping any two adjacent characters in the text if their MI or CF is greater than a threshold. This is equivalent to NN clustering [17] which defines a distance matrix between characters in a clause. Distances between two non-adjacent characters are infinite because overlap grouping is avoided. Thus, it is sufficient to know the distance between adjacent characters. Distances of the same character at the same position must have zero distances. A pre-defined threshold can cut the dendrogram into a sequence of subtrees which represent detected words of the clause.

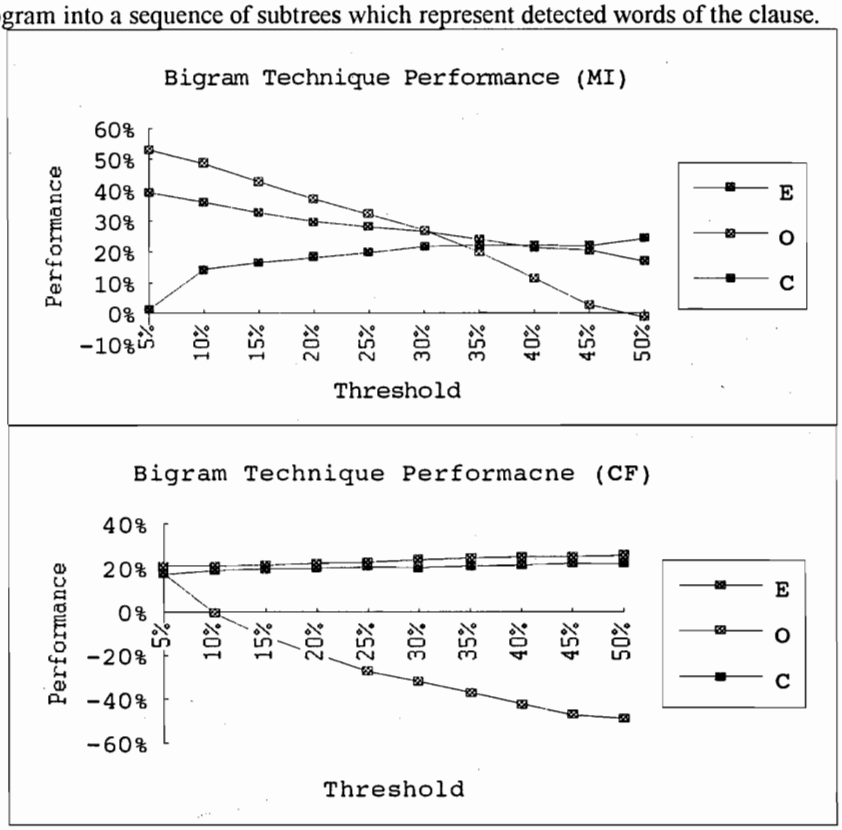


Figure 5: Variation of segmentation performance between distances defined as the MI (a) and CF (b), with respect to different percentage quartiles. Key: E, O and C are the segmentation error, over-segmentation and clause-accuracy, respectively. The suffix m and b indicate the performance measure of MI and the CF, respectively.

The MI is used to define the distance between two adjacent characters and it is estimated from a thematic corpus and not from the a general corpus. The threshold is defined in terms of the top N% quartile since it becomes the N% significance level if the distribution is normal. Apart from MI, the distance can be defined in terms of the CF. Figure 5 shows variation of the segmentation performance between distances defined in terms of the MI and the CF, with respect to different percentage quartiles.

Using MI, the segmentation error reduces steadily from 40% to 17%, just over 50% error reduction. The amount of over-segmentation is about zero when the percentage quartile is about 48% and the clause-accuracy rose from almost zero to about 27%. Note that the accuracy rose dramatically between 5% and 10%. Using CF, the segmentation error and the clause accuracy do not vary dramatically with the percentage quartile. The amount of segmentation error is about the same as the one using MI and the clause-accuracy is only 4% lower than MI. The amount of over-segmentation quickly becomes under-segmentation when the percentage quartile reaches beyond 10%. In summary, the oc-occurrence frequencies are more robust to the

variation of the percentage quartile than using MI but at a cost of lower clause-accuracy and segmentation error.

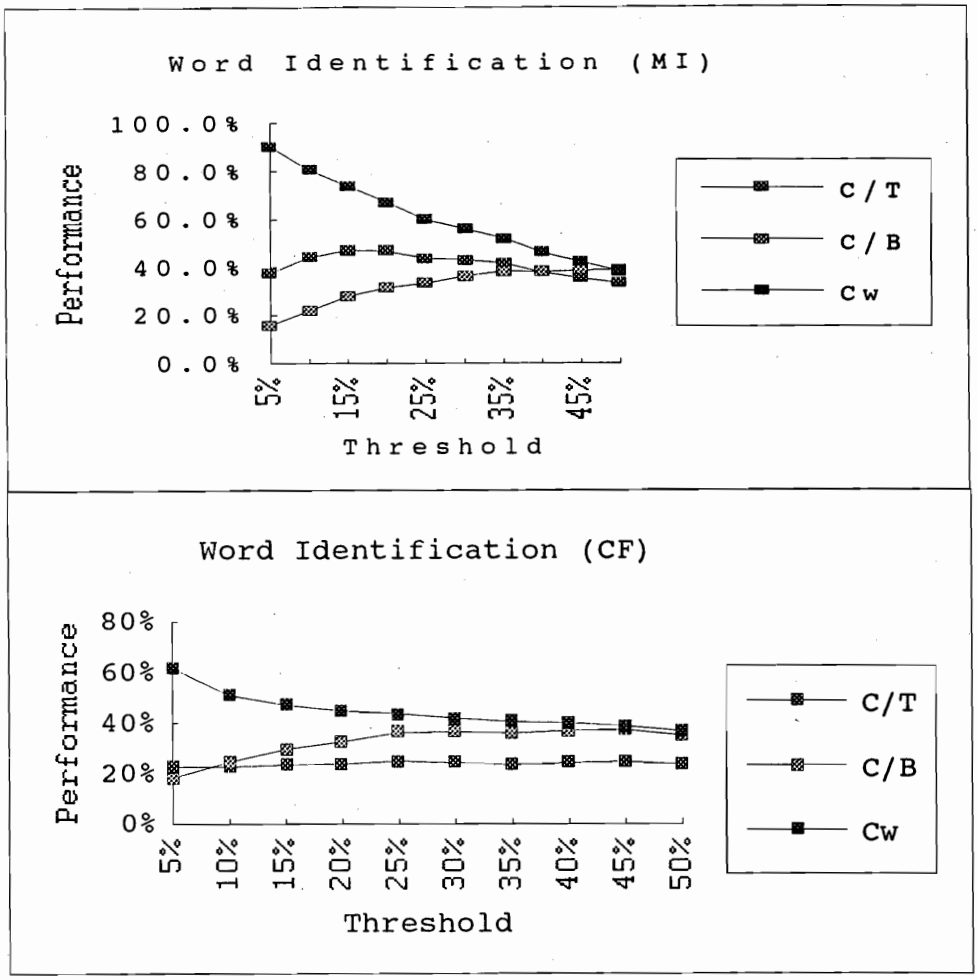


Figure 6: Word-identification performances with different values of percentage quartile for MI (a) and CF (b). Suffixes m and b indicate that NN clustering is carried out using MI and CF, respectively. Key: C/T is the percentage of words in the Basic Law that are identified, C/B is the percentage of identified words in the Basic Law and Cw is the percentage of identified words that are in either the Basic Law or the PH corpus.

Figure 6 shows the word-identification performance with respect to different percentage quartiles. Using MI, the percentages of words in the Basic Law that are identified (i.e. C/T measure) rose to a maximum when the percentage quartile is about 15% where almost 50% of the identified words are words in the Basic Law. A decrease in percentages as the percentage quartile increases imply that there are more words identified but less of them are in the Basic Law. The C/B measures the percentage of identified words that are in the Basic Law. This measure increases steadily but slowly flattened. The Cw measures the percentages of identified words in either the Basic Law or the PH corpus, indicating that the identified words are recognized Chinese words. Here, Cw decreases when the percentage quartile increases, indicating more non-recognized Chinese words are identified.

Using CF, the percentage of words in the Basic Law that are identified do not vary significantly with the percentage quartile. The C/B and Cw measure increase and decrease, respectively, where both asymptotically tend to 36%. Non recognized words (30 characters) are usually longer than those identified using MI (11 characters at the maximum). The word-length distribution of the identified words using bigram frequencies are skewed where as the distribution using MI appears like the distribution of words in the Basic Law. In summary, the CF is more robust than the MI in word-identification but the former can yield almost 100% more correct word-identification than the latter.

The bigram technique achieved similar segmentation performances (i.e. E = 15-17% and C = 28%) to the maximal-matching using words from a general dictionary [16]. The maximal-matching tends to over-segment but the bigram technique can potentially under-segment, depending on the percentage quartile.

V COMBINED TECHNIQUE

The combined technique applies maximal-matching to the text and then using the bigram technique to group single-character words. The identified words are combined with the existing list of words which are used by the maximal-matching to segment the given text again. Figure 7 shows the segmentation performance of the combined technique using MI or CF.

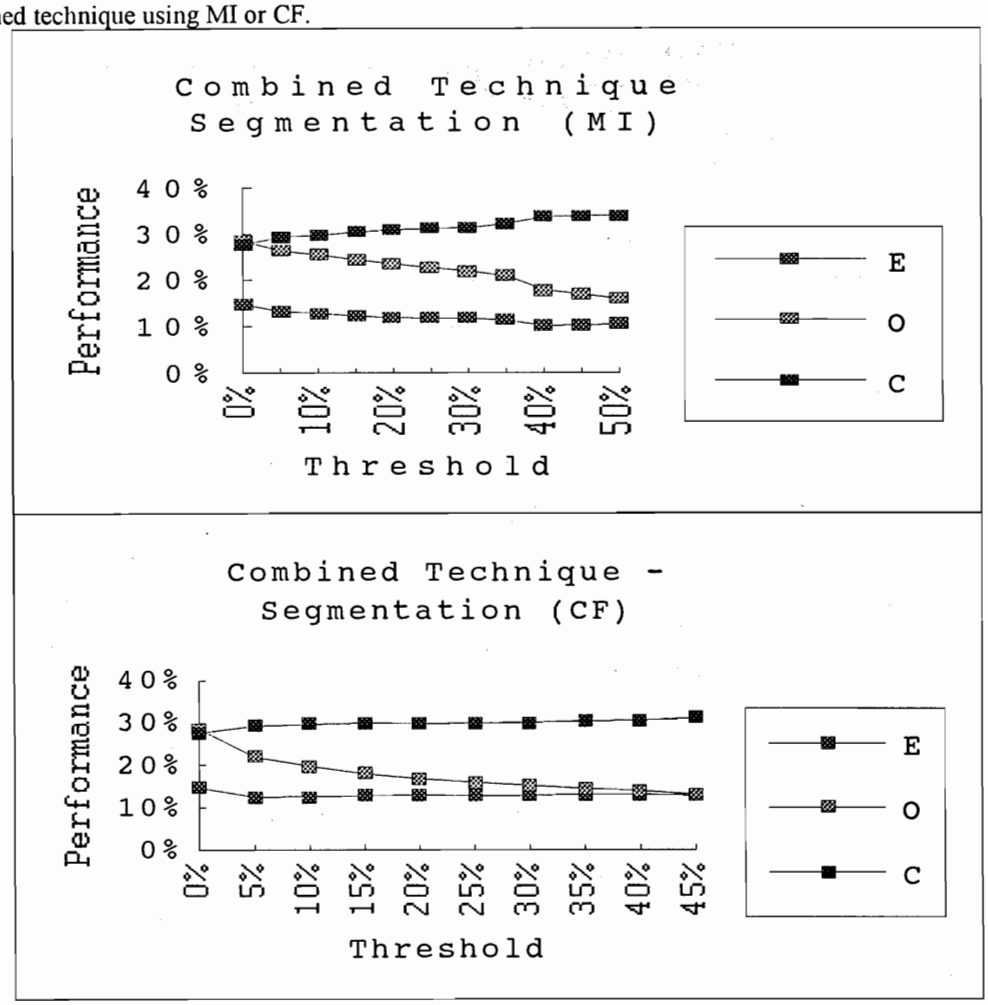


Figure 7: Segmentation performance of the combined technique using. Key: E, O and C are the segmentation error, over-segmentation and clause accuracy, respectively. The following letter "m" and "b" indicate that MI and CF are used, respectively.

Using MI (figure 7), the combined technique can reduce the segmentation error from 15% to 10% (i.e. error reduction of 33%) when the percentage quartile is at 40%. The amount of over-segmentation is reduced by 12% (from 28% to 16%) and the clause accuracy is increased by 5% (i.e. 18% improvement). Using CF, the segmentation error is better than using the MI, only when the percentage quartile is 5%. Otherwise, the segmentation error varies little with the percentage quartile. Reduction of over-segmentation is larger than that using MI but the clause accuracy is not as high as that using MI. In summary, the segmentation performance using MI is better than that using CF.

The amount of correct word identifications are all higher than 50% because the maximal-matching technique uses a dictionary of 52% overlap with the Basic Law (Figure 8). Although the difference between percentages of identified words in the Basic Law is small (C/Tm versus C/Tb) between MI and CF, the other two measures have pronounced difference. In both measures, the MI achieves better word identifications than CF where we expect the percentage of correct identification chosen from the identified words is 80% for

MI, almost independent from the percentage quartile. In addition, the percentage of identified words that are recognized Chinese words remain above 85%, decreasing with increasing values of the percentage quartile. A dramatic increase in word-identification occur in the first 5% quartile and subsequent variation in performance vary less than the first 5%.

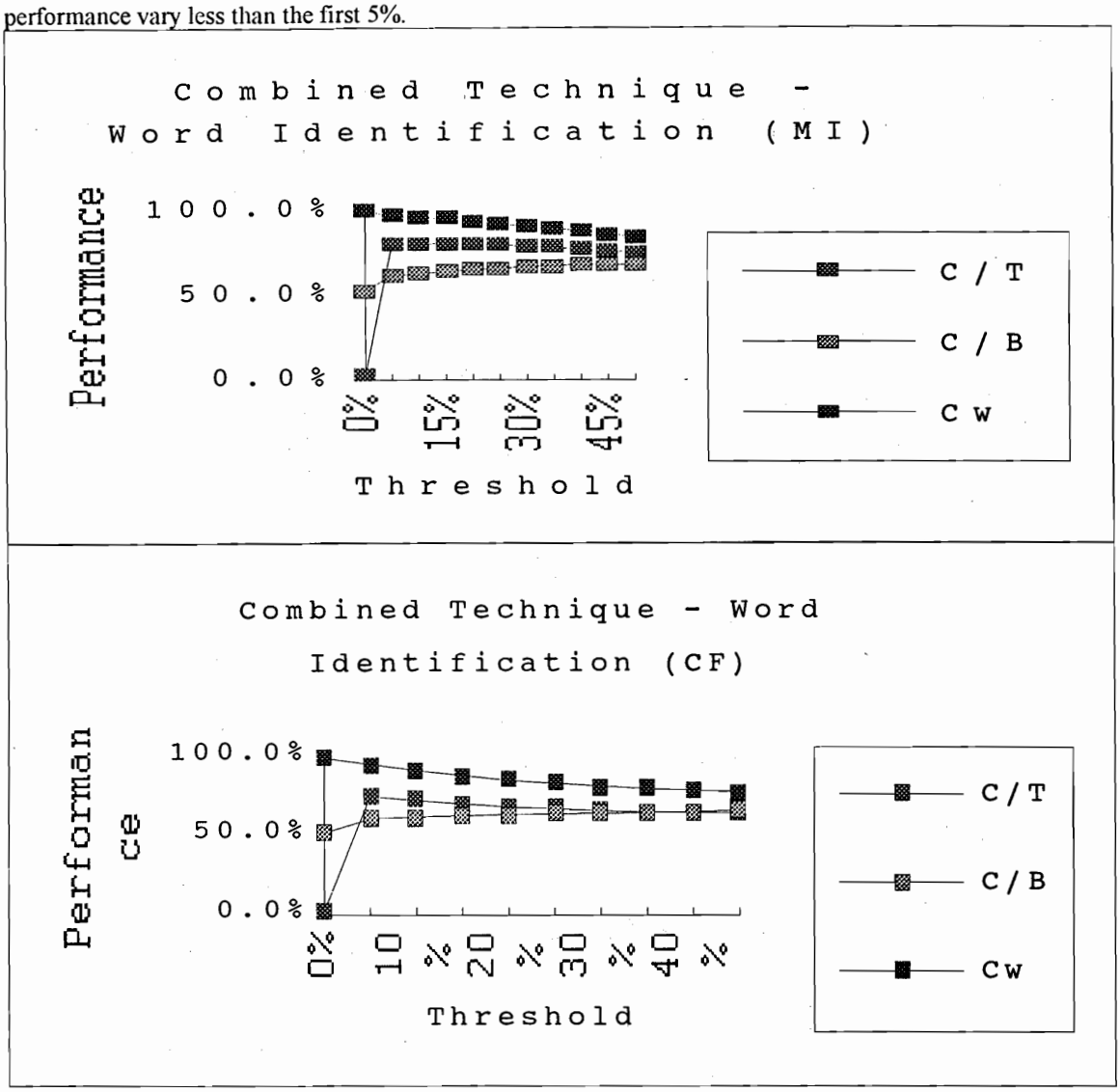


Figure 8: Word-identification performances with respect to different values of the percentage quartile. Key: C/T and C/B represent the percentage of correct identification with respect to the number of identified words and words in the Basic Law, respectively. Cw represents the percentage of identified words that are either in the Basic Law or the PH corpus. The following letter "m" and "b" denote using MI and CF in word segmentation and identification, respectively.

Since the initial list of words used in maximal-matching has 52% overlap with the Basic Law, the measure C/B is not representative. We re-calculated the percentages according to the following normalization formula: dC/T = (C/T - 52%) / (100% -52%). Figure 9 shows that the MI can detect 33% of the remaining words in the Basic Law that are not in the initial word list of maximal-matching. The performance of MI is consistently better than that achieved using CF. When the percentage quartile is 0%, it is equivalent to using only the maximal-matching technique (i.e. no word identification). Again, the first 5% yields a dramatic increase in performance and there is little difference between using MI and CF.

A list of words or short phrases detected by the combined technique is in the appendix where the mutual information is used and the threshold is set at the top 20%. There are no words of length greater than 6. Words of length greater than 4 are few and usually not recognized as words because of the attached verbs (e.g. 屬於). Only 2 out of 17 words of length 4 are not recognized words or phrases. There are more three-

character non-words because of attaching particles of verbs (e.g. 療) or conjunctions (e.g. 及). Function characters at the end of words are not considered to be unrecognized words because they can be detected and rectified. Detection of two character words are more reliable than three-character ones.

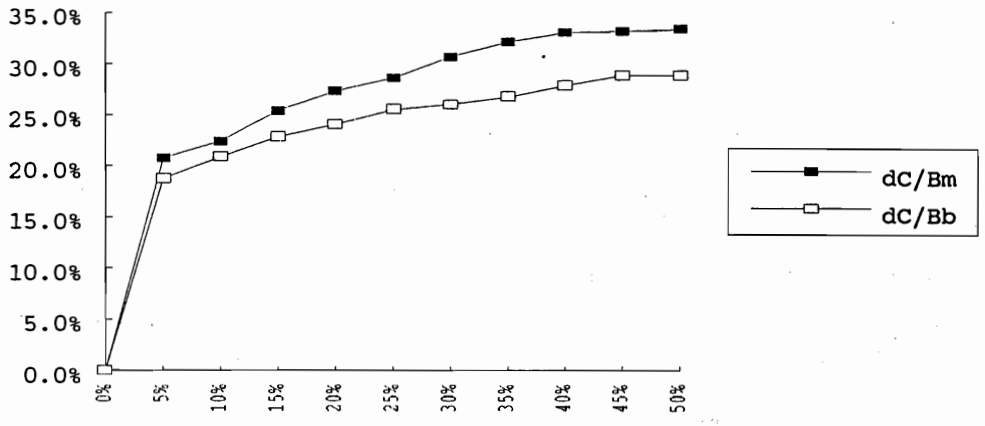


Figure 9: Normalized percentage of identified words that are in the Basic Law but not in the initial word list of maximal-matching. Key: the following letter "m" and "b" denote word segmentation and identification using MI and CF, respectively.

VI. CONCLUDING REMARKS

Our implementation of maximal-matching achieved a mean segmentation error (1.2%) as low as those (1.14%) reported in [11]. However, in practice, when text processing (e.g. machine translation) is applied in a specific domain (i.e. adopt a sub-language approach), the segmentation performance is degraded, as our test data demonstrated (15% segmentation error using words from a general dictionary). The bigram technique which achieves good 2-character word identification offers little assistance to maximal matching as we showed that increasing the amount of 2-character words did not improve the segmentation significantly. We extended the bigram technique to identify words of arbitrary length and the segmentation performance was about the same as maximal matching using a general dictionary. By combining both techniques, we were able to lower the segmentation error by 33% of its degraded performance and improve the word-identification by 33% of the remaining words only in the Basic Law, depending on the percentage quartile (or threshold). The MI appear to yield better segmentation and word-identification performance than CF. However, there is little difference between the two at 5% quartile where the improvement in performance is most dramatic.

REFERENCES

[1] McNaught, J. (1993) "User needs for textual corpora in NLP", *Literary and Linguistic Computing*, **8**, n4, pp.227--234.

[2] Zhang, J-S., S. Chen, Y. Zheng, X-Z. Liu and S-J. Ke (1992) "Automatic recognition of Chinese full name depending on multiple corpus", *Journal of Chinese Information Processing*, **6**, n3, pp. 7-15.

[3] Chang, J.S., C.D. Chen and S.D. Chang (1991) "Chinese word segmentation through constraint satisfaction and statistical optimization", *Proceedings of ROC Computational Linguistics Conference*, Kenting, Taiwan (in Chinese), pp. 147-166.

[4] Guo, J. and H.C. Lam (1992) "PH: a Chinese corpus for pinyin-hanzi transcription", *Technical Report TR93-112-0*, Institute of Systems Sciences, National University of Singapore.

[5] Smadja, F. (1993) "Retrieving collocations from text: Xtract", *Computational Linguistics*, **19**, n1, pp.143--177.

[6] Salton, G. (1989) *Automatic Text Processing*, Addison-Wesley: Reading, Mass.

[7] Sproat, R. and C.L. Shih (1990) "A statistical method for finding word boundaries in Chinese text", *Computer Processing of Chinese and Oriental Languages*, **4**, n4, pp. 336-351.

[8] Su, M.S. (1993) *Private Communication*, Tsinghua University, Beijing, People's Republic of China.

[9] Kit, C., Y. Liu and N. Liang (1989) "On methods of Chinese automatic word segmentation", *Journal of Chinese Information Processing*, **3**, n1, pp. 13-20.

[10] Fan, C.K. and W.H. Tsai (1988) "Automatic word identification in Chinese sentences by the relaxation technique", *Computer Processing of Chinese and Oriental Languages*, **4**, n1, pp. 33-56.

[11] Chiang, T.H., J.S. Chang, M.Y. Lim and K.Y. Su (1993) "Statistical models for word segmentation and unknown word resolution", *Proceedings in ROCLING V '93*, pp. 123-146.

[12] Zu, B.Z., J. Zhang and Q.H. He (1992) "The method of Chinese word segmentation based on neural network", *Journal of Chinese Information Processing*, **7**, n2, pp. 36-44.

[13] PRC Governement Publication, *Hong Kong Basic Law*.

[14] Lun, C.S. and L. Ling (1993) *Hong Kong Basic Law (in Big-5 Code)*, Department of Chinese, Translation and Linguistics, City Polytechnic of Hong Kong.

[15] National Standard Bureau (1988) *Contemporary Chinese language words segmentation standard used for information processing*.

[16] Fu, X-L. (1987) *Xiandiao Hanyu Tunrun Cidian*, Waiyu Jiaoxue Yu Yanjiu Publishing House: Beijing, PRC.

[17] Everitt, B. (1985) *Cluster Analysis*, Heinemann: London.

APPENDIX

The following is the list of words detected by the combined technique from the Hong Kong Basic Law. The threshold is set at the top 20% of all bigrams measured by mutual information. If the last character of an entry is a slash character (i.e. "/"), then the entry is a plausible word or short phrase. Due to space, 2-character words detected are not included here.

應課差餉租值/	不少於/	這幾個/
自然資源屬於	紫荊花/	白兩色
衡又互相配合	不低於/	刑事罪/
違反誓言而	療衛生	獲功能/
各類院校均	審計署/	學語言/
集裝箱碼頭/	及竊取	來源證/
過半數票即	準備金/	境衛生
專題小組/	約束力/	航空器/
出口配額/	屈功能	新興產/
收支平衡/	丁屋地/	將採取/
過半數票/	盡忠職	也先後/
醫療衛生/	彈劾案/	範圍及
開支標準/	星花蕊	登記冊/
廉潔奉公/	範圍內/	記錄在/
日恢復對	或瀆職	明創造
技術停降/	證明書/	標準向
鄉村屋地/	代擬稿/	興旺發
軍用船隻/	既互相	均假定/
自然資源/	原舊批	花蕊上/
顆象徵著	過半數/	西醫藥/
刑事罪犯/	姬鵬飛/	龍半島
登記標誌/	構想及	
外圈寫有/	元匯價/	
救助災害/	製造業/	
航空公	學歷等/	
預算案/	被判犯/	

The Acquisition and Expansion of Knowledge Data By Analyzing Natural Language
−Using Five-Character Kanji (Chinese character) strings−

YASUHITO TANAKA
Aichi Shukutoku University
Aichi−Shukutoku University 9 Katahira Nagakute Nagakute Aichi 〒480−11 Japan
TEL +81-561-62-4111 FAX +81-561-62-3007

Abstract

Knowledge data are indispensable for the comprehension and context analysis of natural language. The author describes the ways of acquiring and expanding such knowledge data. Kanji (Chinese character) strings are frequently used in the Japanese language. The author attached importance to five-character Kanji strings and decided to extract the five-character strings which can be divided into two character ⊕ three-character or three character ⊕ two-character combinations. A large quantity of such data were collected and knowledge data were further expanded by combining them with postpositive particles and auxiliary verbs.

Five-character strings were extracted from the Asahi Shimbun, and about 76,000 items of knowledge data were obtained by sorting them out. The knowledge data thus obtained could be further expanded.

1. Introduction

Knowledge data are necessary for the comprehension and context analysis of natural language. How can a large quantity of such knowledge data be prepared?

The methods employed in, this study to collect and expand such knowledge data are described below. To begin with, the method of collecting the data will be described. Five-character strings were used in collecting the data, and they were collected from one-year old issues of the Asahi Shimbun.

2. Why were five-character strings selected?

Five-character strings were chosen for the study for the following reasons.

(1) Five-character strings could be selected mechanically, and five-character strings are numerous following four-character strings. The number of two-character strings is the largest among all types of character strings.

(2) Five-character strings can be divided into two-character ⊕ three-character or three-character ⊕ two-character combinations, or combined words. Two-character and, three-character are basic terms and occur very frequently.

(3) The category of basic terms can be expanded into phrases with the addition of particles and other words inserted between the two component parts.

Example 1.

経営・多角化 Keiei takakuka
↓
経営を多角化する Keiei wo takakukasuru (to diversify management)
経営の多角化 Keiei no takakuka (diversification of management)

Example 2.

自主的・判断 Jishuteki handan
↓
自主的な判断 Jishutekina handan (autonomous judgment)
自主的に判断する Jishutekini handansuru (to judge autonomously)

Some strings form phrases by reversing the order of the two parts.

Example 3.

自動車輸出	Jidosha yushutsu
↓	(automobile export)
自動車を輸出する	Jidosha wo yushutsusuru
	(to export automobiles)
自動車の輸出	Jidosha no yushutsu
	(export of automobiles)
輸出した自動車	Yushutsushita jidosha
	(exported automobiles)
輸出する自動車	Yushutsusuru jidosha
	(automobiles to be exported)

(4) It is necessary to translate the same term in a variety of ways. Therefore, it may be studied as a technical term or as a translati on selected for machine translation.

(5) It is easy to expand knowledge by utilizing two-character ⊕ three-character or three-character ⊕ two-character combinations collected from five-character kanji strings. Detailed explanations will be made in this article.

3. Acquisition of knowledge data through partition of five-character kanji strings

3.1 Collection of five-character kanji strings

Five-character kanji strings can be extracted mechanically from Japanese text.

Five-character kanji strings can be classified into two types of combinations of basic words; namely, a two-character ⊕ three-character combination or a three-character ⊕ two-character combination. The two component words can be changed into sentences or phrases.

In this study, 76,000 different five-character kanji strings were extracted from data contained in one-year old issues of the Asahi Shimbun. T he total number of five-character Chinese character strings was 210,000.

The number of different five-character kanji strings divided by the total number of five-character Chinese character strings that occurred was 0.36. Of these, 39,000 strings were usable as knowledge data of concurrence relations.

The data were sorted out according to the following procedures.

1) Extraction of five-character kanji strings from Corpus

2) Compression of same-character strings and analysis of their frequency

3) Reference to already sorted out knowledge data (This procedure was omitted this time.)

4) Examination of content and storage as knowledge data

A book explaining the content of these procedures is scheduled to be published in the near future. Provided in the book will be data classified in the order of prepositional and postpositive particles.

Classification code

23…Data classifiable into 2 character ⊕ three character combinations
(仏人研究者) (Futsujin kenkyusha)
(French researcher)

32…Data classifiable into three character ⊕ two character combinations
(日本人気質) (Nihonjin katagi)
(Japanese trait)

70…Those classifiable into other than 23 and 32
(悲観主義者) (Hikan shugisha)
(pessimist)

80…Names of persons, enterprises and other proper names
(～市議会、～営業所、～役場)
(～shigikai (municipal assembly),
～eigyosho (business establishment),
～yakuba (local government office))

90…Place names

99…(unintelligible strings, strings requiring explanatory particles)
(同日朝現在、連勝中中国)
(Dojitsu asa genzai, renshochu chugoku)

Classification code	Types	Total number of combinations
23	17,705	57,271
32	22,076	60,092
70	8,823	24,316
80	10,101	41,988
90	1,244	2,744
99	16,293	25,381
Total	76,242	211,793

Data extracted from newspapers are characterized by a relatively high frequency of

names of persons, enterprises and places.

A further examination of data classified into the 99 category will serve to improve the technique of collecting terms and analyzing them into form elements.

3.2 Analysis of component words

1) Results of analysis of five-character kanji strings classifiable into two-character ⊕ three-character combinations

	2-character strings	3-character strings
Code 23	4,199	7,467

2) Results of analysis of five-character strings classifiable into three-character ⊕ two-character combinations

	3-character strings	2-character strings
Code 32	8,253	3,914

3) Totals of two-character words and three-character strings

Two-character strings

4,199 + 3,914 = 8,113 ⇒ 6,417

Three-character strings

(duplications deleted)

7,467 + 8,253 = 15,720 ⇒ 13,527

These two-character and three-character string terms are basic and numerous in occurrence.

In the analysis of five-character strings, the reader's comprehension will be facilitated if the strings are written with a space between the component words. However, since all of these strings cannot be processed by machine, human interference is required.

If five-character strings are classified according to three-character and two-character component words, the following four combinations are possible.

1	○	○
2	○	×
3	×	○
4	×	×

(However, three-character words are processed on a priority basis.)

It is necessary to classify data largely in this way before it is examined by humans.

3.3 Simplification of coding work

Many five-character strings can be classified into three-character + two-character and two-character + three-character combinations. If we extract only three-character words, and analyze the final characters that appear within them, special characters can be extracted.

Therefore, it is possible to classify five-character strings into three-character ⊕ two-character and two-character ⊕ three-character combinations quickly by mechanically extracting five-character strings and analyzing the data by utilizing these words.

Examples : ～ teki, ～ ka, ～ sha, ～ sho

～的、 ～化、 ～者、 ～所

The following are three-character kanji strings in which teki and ka appear as the final characters.

～化			～的		
	種類	延件数		種類	延件数
民営化	35	421	政治的	195	580
自由化	33	93	国際的	115	344
合理化	23	102	具体的	110	409
近代化	19	58	社会的	106	362
民主化	16	78	経済的	92	454
情報化	14	93	歴史的	85	240
国際化	14	69	基本的	78	384
工業化	13	21	軍事的	53	126
実用化	10	17	本格的	49	83
活性化	10	13	技術的	47	76
一本化	9	22	個人的	42	98
機械化	9	12	世界的	42	80
自動化	9	9	戦略的	42	67
商品化	7	8	精神的	41	103
高齢化	6	229	国民的	39	169
軍事化	6	10	積極的	38	104
多角的	5	15	科学的	36	82
砂漠化	5	14	一方的	32	85
国産化	4	20	代表的	29	60
空洞化	4	16	比較的	29	41
保守化	4	16	総合的	28	45
孤立化	4	7	心理的	27	57
省力化	4	7	国家的	25	44

Furthermore, coding work can be simplified by classifying five character kanji strings in the following way.

(1) X X X X X — Effective for analyzing three-character ⊕ two-character combinations
2 1 3
Order of classification

(2) X X X X X — Effective for analyzing three-character ⊕ two-character and two-character ⊕three-character combinations
4 3 2 1
Order of classification

(3) X X X X X — Effective for analyzing three-character ⊕ two-character and two-character ⊕three-character combinations
1
Order of classification

It is necessary to begin the coding task with the component that can be coded most easily, and to change the order of classification so that the coding can be accomplished accurately and quickly.

3.4 Detailed coding work

Data falling in the 70 category can be divided among 14 different detailed classifications. (For the purpose of eliminating 2, 3 ; 3, 2) However, the number of classifications becomes larger if the connection relations and parallel relations are considered in detail.

Partition patterns of five-character kanji strings

		Seq. No.
1)	5	1
2)	1, 4	2
	4, 1	3
3)	3, 2 ⎫ Coded	4
	2, 3 ⎭	5
	3, 1, 1	6
	1, 3, 1	7
	1, 1, 3	8
4)	2, 2, 1	9
	2, 1, 2	10
	1, 2, 2	11
	1, 1, 1, 2	12
	1, 1, 2, 1	13
	1, 2, 1, 1	14
	2, 1, 1, 1	15
5)	1, 1, 1, 1, 1	16

16 patterns of partitions

Generally speaking, the number of partitions is few in five-character strings, and in many cases the number of characters in partitioned words is the same. This is considered to be due to a phenomenon of optimization of the labeling of concepts or word expressions.

On this basis, it is necessary to perform detailed coding work on about 8,800 items of data coded in the 70 category.

4. Expansion Method of Knowledge Data

4.1 Word-to-word relations in text

If we examine how five-character kanji strings are arranged in Japanese text, we will see what verbs and auxiliary verbs make up the phrases contained in the basic conceptual words of two-character ⊕ three-character and three-character ⊕ two-character combinations. This can be seen by analyzing data into phrases. It is also possible to examine this by forming actual sentences by making KWIC.

● 1 character ●、● 2 characters ●
● 3 characters ●、● 4 characters ●
● 5 characters ●、……

Here, ● represents a component element of five-character kanji strings.

It is possible to use the above relationship to make KWIC and collect new data in a concentrated way.

As for word-to-word relations and connectives, refer to Table 1.

4.2 Expansion of knowledge data

An attempt is made to expand knowledge data on the basis of word-to-word relations and connectives.

Example:

		With/without relation
新幹線建設	Shinkansen kensetsu	
↓		
新幹線の建設	Shinkansen no kensetsu	O
新幹線が建設	Shinkansen ga kensetsu	X
新幹線を建設	Shinkansen wo kensetsu	O
新幹線に建設	Shinkansen ni kensetsu	X
新幹線で建設	Shinkansen de kensetsu	X

Only correct relations are marked and extracted.

● ・ ●

If the preceding word is a noun forming a Sa-hen verb,

● する ●　　● suru ●
● した ●　　● shita ●

are formed.

Example:

情報化・社会→情報化した社会

Johoka shakai → informationalized society

Furthermore, if it is the root of an adjective, −teki na is added to form the following.

● 的な ●　　● tekina ●
● 敵に ● する　　● tekini ● suru

When ● is a Sa-hen verb.

Example:

実質・成長率→実質的な成長率

Jisshitsu seichoritsu → Jisshitsuteki na seichoritsu (real-term growth rate)

However, such combinations are required to develop the components into phrases for examination.

As there are cases in which the component words of two-character ⊕ three-character and three-character ⊕ two-character combinations are used independently, it is necessary to see if such component words are found in the dictionary and if they are not, it is necessary to add them.

In this way it is possible to collect five-character strings exhaustively, without any omission, and to increase knowledge data with some assurance. Furthermore, it is also possible to obtain printouts and distribute them to a large number of people, and as a result, to obtain and expand knowledge data.

There are connections with special verbs between two base words. It is all right if such connections are input as proper base-word-base word connections.

5. Other Methods of Expanding Knowledge Data

5.1 Expansion of knowledge by association

所得税・減税　　Shotokuzei • genzei
↓
所得税を減税する　　Shotokuzei wo genzeisuru
↓　　(to reduce income tax)

(i)所得税を減らす　(i) Shotokuzei wo herasu (to reduce income tax)
(ii)税金を減らす　(ii) Zeikin wo herasu (to reduce tax)

It is possible to expand knowledge by utilizing a thesaurus and synonyms.

5.2 It is also important to reverse the order of two-word combinations and examine if there are relations between the reversed words.

Example:

貿易・不均等　　貿易が不均等
Boeki • fukinko　Boeki ga fukinko (trade imbalance)

不均等・貿易　　不均等な貿易
Fukinko • boek　Fukinkona boeki (unbalanced trade)

Knowledge data can be expanded by utilizing the existing 390,000 items of knowledge data. All word combinations are not reversible.

5.3 Generation of compound words

Two-word combinations have the following relations with one word as the center.

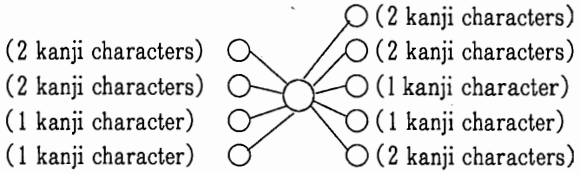


Fig. 1 Compositions of words

It is possible to examine the generation conditions of compound words by producing a composition diagram such as the one shown above.

6. Translation of Compound Words

6.1 English translations of five-character kanji strings

Some five-character kanji strings were selected and given English translations. The following became clear from these translations.

(i) English translations had diverse meanings

(ii) The joining of translated words is as follows.

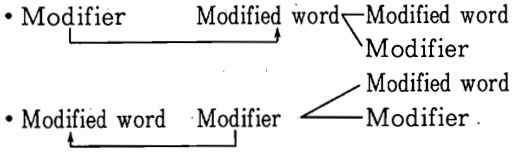
Translation of base word ⓧ translation

of base word－resulting translation

Noun		noun		noun
Verb	ⓧ	verb	→	verb
adjective		adjective		adjective
adjectival ver		badjectival ver		badjectival verb

(iii) Translation of compound words

The translations of compound words differ according to whether they are modifiers or modified words.



(iv) Others

- Japanese base words do not necessarily correspond to English words word by word.
- This question is related very much to the quality of translators. Attempts should be made to automate or semiautomate translation.
- Translation work is costly, so it is desirable to reduce translation costs.
- In view of the large quantities of translation work needed, it is necessary to develop a system to improve the speed and ensure the accuracy of translation.

6.2 Translation of ordinary compound words

Translation of compound words is an important task. While it is necessary to use generally accepted translations of technical terms, etc., it is also necessary, after considering the details described in 6.1, to select translations according to the following procedures.

(i) To partition compound words into base words

(ii) To relate base words and partition them structurally

(iii) To give proper translations

This procedure is shown in the following example.

自然言語処理…複合語 Shizen gengo shori…Compound word

↓

自然・言語・処理…単語分割 Shizen • gengo • shori…Partition into words

↓

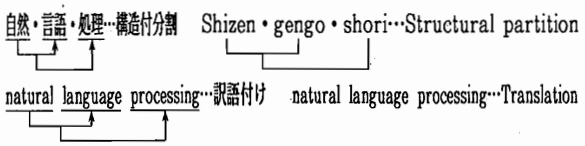


Fig. 2 Process of translation

In order to complete a system of this kind, it is necessary to devise a system by which corresponding translations to base words are collected and selected for combination.

7. Reference to Thesaurus

A thesaurus collects and sorts words according to similar concepts and systematically develops them into upper-ranking concepts.

As words are grouped according to similar concepts, it is possible to systematically grasp the characteristics of words. With the aid of other reference sources, the thesaurus serves to resolve the problems presented by the diverse meanings of the same words.

Word-to-word relations in this study should be referred to the Thesaurus and their concepts should be analyzed in detail. The thesaurus plays an important role in this respect.

7.1 Reference to Thesaurus

The reference of hundreds of thousands of items of knowledge data to the Thesaurus makes, it possible to supplement deficient knowledge data and extend the classification of concepts in the Thesaurus according to the meanings of words. It also enables machine translation to select accurate translations for words with different meanings. It is expected that a large scale machine-readable thesaurus will be provided.

Word-to-word relations can be obtained by the partition of five-character kanji strings and through additions of verbs or auxiliary verbs to them. The number of word-to-word relations will be tremendously large, and the task of collecting them and sorting them will pose a difficult problem.

If we consider a number of words (n), combinations multiplied by itself (n2) are possible, and with the addition of verbs, auxiliary verbs

and other words, the number is increased to kn2. Very few of these Kn2 combinations make sense, but they must be examined.

Since this involves a tremendous amount of work, it may be possible to refer to the Thesaurus system and focus on the examination of groups of combinations which are supposed to make sense.

In this way, it is possible to keep a nearly infinite number of relations down to a limited number.

1) Meaning of reference to Thesaurus (i)

委員長就任 → 委員長が就任 Iincho shunin → Iincho ga shunin
　　↘ 委員長を就任 　　↘ Iincho wo shunin

In this word-to-word relation, are Iincho (chairman), gicho (chairman of a meeting), kaicho (board chairman), shacho (president), officers, etc., on the same semantic marker?

To confirm the expansion of knowledge data and expanded semantic markers through reference to the Thesaurus

2) Significance of reference to Thesaurus (ii)

Reference of word-to-word concurrence relations of five-character kanji strings to the Thesaurus serves the following purposes.

- To facilitate the verification of the accuracy of the Thesaurus
- To facilitate the subdivision and integration of semantic markers
- To facilitate judgment in the selection of separate translations and in the extraction of exceptions

3) Meaning of reference to Thesaurus (iii)

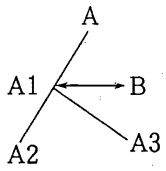
By combining knowledge data on word-to-word relations in five-character kanji strings with the Thesaurus, it is possible to know with what concepts verbs are combined. If it is known that there is a connection relation between A1 and the verb B, it is possible to examine whether all the words in Group A2 to which A1 belongs can be combined with the verb B.

If the connection with A1 makes sense, it may be considered a combination of A1 and B.

If it is known that the words in Group A2 can be combined with B, it is necessary to examine whether the same translation can be used.

Furthermore, the connection is expanded to A3, and its connection with B is examined. Similarly, connections are further developed to an upper-ranking concept of the Thesaurus. It is possible to save a tremendous amount of labor in this way.

It is because of this, also, that prepositional words are classified as basic conceptual words in knowledge data on word-to-word relations. The partitioning of five-character kanji strings is of great significance for this purpose, too.



7.2 Thesaurus and long-unit words

Compound words and long-unit words are used very frequently in text. There is a way of extracting base words from long-unit words, but for this purpose, it is necessary for many long-unit words to be incorporated in the Thesaurus.

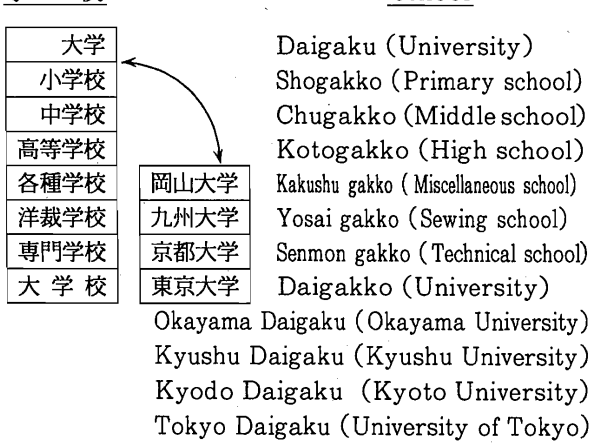


Fig. 4 Groups of words with lower level meanings are studied in order to further develop word-to-word relations, and upper and lower ranking groups are formed in relation to groups of words with lower level meanings.

7.3 Thesaurus and systematization of knowledge

In order to realize a high level machine translation system and sentence comprehension system, it

is necessary to input knowledge in machines. This system of knowledge is represented by a thesaurus system and a system of concepts.

This system is illustrated as follows.

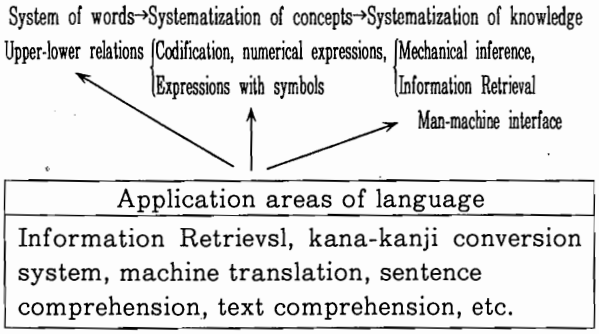


Fig. 5 Thesaurus of systematization of knowledge

It may be said that we use the Thesaurus to evolve a system of knowledge for incorporation into machines.

8. Evaluation of Knowledge Data

A method has been established that enables us to collect large quantities of knowledge data. In the future it will be necessary for us to evaluate the knowledge data, i.e., to examine what data remains to be collected, to learn the extent to which the collected data is duplicated, and to find the answers to other questions.

It is also important to create an environment that permits additions and revisions to the collected knowledge data.

A step has just been taken toward the work of extracting large quantities of knowledge data. In order to incorporate knowledge data in machine translation systems, it is necessary for us to go through the following stages.

It is also vital to establish a method for evaluating a thesaurus, and to specify the parameters of a thesaurus.

The following can be suggested as the parameters of a thesaurus.

1) No. of words in the thesaurus
2) Application areas of the thesaurus
3) Contents of the thesaurus and its machine r eadability
4) Provision of various kinds of utilities for the use of the thesaurus
5) Relations between the succession of knowledge and inference systems, and other items.

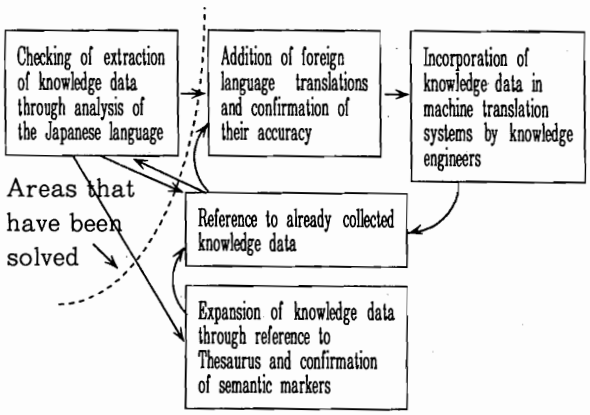


Fig. 6 Work processes until knowledge data on word-to-word relations are incorporated within a machine translation system

9. Future Tasks

1) To translate all knowledge data in five-character kanji strings

- Translation and checking cost per case ¥500
- Cost for translating 39,000 cases ¥19.5 million
- Work volume per day (1 person) 100 cases/day
- Total number of man-days (about one year by 2 persons) 390 man-days

If the budget is limited, it is possible to begin the task with words that occur with a higher frequency, or to begin with a certain word and continue the work sequentially. Part of the contents of translations are to be shown upon completion of the task.

If the these expenses are to be paid, it is necessary to consider a proper distribution of costs to excellent translators and the costs for the examination and reference of data.

Furthermore, it is necessary to think of ways to improve speed and maintain the accuracy of translations.

2) Regarding data which are not collected in this experiment, it is necessary to study whether the lack of such data is due to the differences

of the areas covered or to the obsoleteness of the data itself, and other related matters.

3) It is necessary to make efforts to practically apply knowledge data on word-to-word relations to a machine translation system, kana-kanji conversion system, and voice and character recognition systems.

4) If large quantities of knowledge data on word-to-word relations in five-character kanji strings are obtained at low cost, studies on natural language will necessarily develop in a new direction. Just as it is said in philosophy that quantitative expansion leads to a qualitative change, studies on natural language must move onward to a new stage.

We are now in an era in which large quantities of knowledge data on word-to-word relations can be obtained at low cost.

The provision of large quantities of knowledge data

1) can systematize grammar (simplification, sophistication of grammar),
2) prevent the generation of a large number of structures in sentence structure analysis,
3) resolve the existence of many meanings of the same words in machine translation and improve the results of machine translations,
4) serve to improve the precision of character and voice recognition,
5) simplify the differentiation of homonyms and words of same the forms with different meaning, and
6) promote the development of semantic analysis in natural language processing.

10. Conclusion

It is desirable that knowledge on natural language be acquired automatically, but this poses difficult problems, due partly to the fact that knowledge data are needed in order to resolve the problem of the different meanings that exist for the same words in sentence structure analysis.

In this article, the author discussed five-character kanji strings obtained in a simple way, and methods of sorting them and expanding them. In addition, a study was made on giving translations which are necessary for machine translation.

The author is indebted to Yoriko Inada for her cooperation in the data-sorting work in this study, to the staff members of the Aoe Research Laboratory at the University of Tokushima for their cooperation in extracting five-character kanji strings and inputting them in code, and finally to Mr. Yoshikawa and Mr. Kato of Brother Industries, Inc., for their cooperation in data analysis.

References:

(1) Yasuhito Tanaka and Masaru Yoshida: Acquisition of Knowledge Data by Analyzing Natural Language, 11th International Conference on Computational Linguistics COLING, '86, August 1986
(2) Yasuhito Tanaka and Masaru Yoshida: Knowledge Data (Word-to-Word Relations) and solution of Multivocal word, Natural Language Processing, Information Processing Society, 60-3, March 1987
(3) Yasuhito Tanaka: Data for Analysis of Word-to-Word Relations – Explanations and Materials with "wo" as the Center (1), (II)
The Summing-Up Group on "Sophistication of Language Information," a designated scientific research project subsidized by the Ministry of Education, March 1987
(4) Yasuhito Tanaka: On Knowledge Data Based on Word-to-Word relations, "Quantitative Linguistics," Collection of Articles on the Japanese Language, Akiyama Shoten, March 1987

Table 1
Method of Acquiring Knowledge Data
(Method of expanding knowledge within frameworks, acquisition of knowledge through verification)

1－1	～の～	～no～
1－2	～を～	～wo～
1－3	～が～	～ga～
1－4	～な～	～na～
1－5	～に～	～ni～
1－6	～で～	～de～
1－7	～や～	～ya～
1－8	～も～	～mo～
1－9	～と～	～to～
1－10	～へ～	～e～
1－11	～的～	～teki～
1－12	～性～	～sei～
1－13	～化～	～ka～
1－14	～・～	～・～
1－15	～，～	～，～
2－1	～から～	～kara～
2－2	～され～	～sare～
2－3	～した～	～shita～
2－4	～する～	～suru～
2－5	～との～	～tono～
2－6	～とかへ～	～toka～
2－7	～ない～	～nai～
2－8	～なり～	～nari～
2－9	～にも～	～nimo～
2－10	～には～	～niha～
2－11	～まで～	～made～
2－12	～への～	～eno～
2－13	～より～	～yori～
2－14	～での～	～deno～
2－15	～では～	～deha～
2－16	～的に～	～tekini～
2－17	～的な～	～tekina～
2－18	～性の～	～seino～
2－19	～性を～	～seiwo～
2－20	～上の～	～jono～
2－21	～側に～	～gawani～
2－22	～化の～	～kano～
2－23	～内の～	～naino～
3－1	～および～	～oyobi～
3－2	～された～	～sareta～
3－3	～される～	～sareru～
3－4	～しうる～	～shiuru～
3－5	～すべき～	～subeki～
3－6	～すると～	～suruto～
3－7	～だけを～	～dakewo～
3－8	～ている～	～teiru～
3－9	～できる～	～dekiru～
3－10	～である～	～dearu～
3－11	～という～	～toiu～
3－12	～として～	～toshite～
3－13	～と同じ～	～to onaji～
3－14	～とかの～	～tokano～