A COMPUTER SYSTEM FOR IDENTIFICATION AND STATISTICAL ANALYSIS OF SYLLABLES IN THE ROMANIZED ZHUANG SCRIPT

Tom B. Y. Lai and Chun-yu Kit
Department of Applied Linguistics
City Polytechnic of Hong Kong
83 Tat Chee Avenue, Kowloon, Hong Kong
E-mail:ALTOMLAI@CPHKVX.BITNET

Abstract

The Computer System for Identification and Statistical Analysis of Syllables in the Romanized Zhuang Script - CISS-Zhuang - was developed for use in a research project, now being undertaken at City Polytechnic of Hong Kong, that aims at studying the relationship between Zhuang and Chinese, in particular Cantonese, lexicons.

The Zhuang language is a minority language that is spoken in southwest China. In order to compare lexical items in Zhuang and Chinese, a dictionary of Zhuang has been input into a computer data base. The Zhuang items, in their romanized spellings, also serve as a key to collating smaller data bases of findings of research on the relationship between Zhuang and Chinese.

The CISS-Zhuang system was designed and implemented on a PC/AT compatible computer. Techniques used in Chinese automatic word segmentation (or identification) were borrowed because the task was of a similar nature. A small knowledge base has been built for disambiguation in the process of syllable identification and for detection of spelling mistakes.

I. Introduction

I.1 The Project

The Computer System for Identification and Statistical Analysis of Syllables in the Romanized Zhuang Script - CISS-Zhuang - was developed for use in a research project, now being undertaken at the City Polytechnic of Hong Kong, that aims at studying the relationship between Zhuang and Chinese, in particular

Cantonese, with an emphasis on the lexical asspects.

Zhuang is the language of a minority people of over 13 million in China. [1] It is closely related to other minority languages spoken in China like Puyi and Dai, and outside China, Zhuang has close relatives like Thai, Laos and Shan in countries including Thailand, Laos, Burma, Vietnam and India. Zhuang and its close relatives are thus spoken in an extensive area comprising Southwestern China and a large part of the Indo-China Peninsula. In History, this non-Sinic language was believed to have been spoken in the whole of southern China from the China Sea coast in the east to the high mountain ranges in the west. As most of these parts of China are now inhabited by Chinese-speaking people, it is not unreasonable that the Zhuang group of languages once spoken by the original inhabitants should have left their marks in the form of sub-stratum elements.

The Cantonese dialect of Chinese, spoken in the provinces of Guangdong and Guangxi, which are within the geographical area inhabited by ancient speakers of languages of the Zhuang group, is thus a language laden with sub-stratum Zhuang elements. [2] The authors have begun a study of the Zhuang lexicon with a view to compare lexical items in Zhuang and Cantonese Chinese. A dictionary of the "standard" Zhuang language has been input into a computer data base. Now, cognates in the two languages as well as loanwords in both directions as identified in previous linguistic studies are ready to be input into the database so that findings of different linguists can be compared and collated. As an initial effort, the authors have converted relevant findings in four papers into data bases that are ready to be collated with the dictionary data base. Some preliminary observations of the authors have also been input into a database.[3]

I.2 The Role of Syllable Indentification

The 22,985 entries of a bilingual dictionary of Zhuang (meanings given in Chinese) published in the Zhuang autonomous region (province) of Guangxi (1984) were input into a computer data base with the the form of the individual words (dictionary entries) as keys. These words are in spelt in the standard writing system designed by the Chinese authorities for writing the Zhuang language. This romanized script will be described in a later section. Here, it suffices to say that it is an artificially designed system, complete

with well-defined spelling and tone (Zhuang is a tonal language) representation conventions as well as rules governing syllable structure and demarcation. Using these rules, CISS has been developed to segment poly-syllabic words in the data base into their constituent syllables. Application of this system yields statistical information about the Zhuang lexicon like the following:[4]

Number of monosyllabic words: 8470
Number of bisyllablic words: 12556
Number of trisyllabic words: 1871
Number of four-syllable words: 80
Number of five-syllable words: 3
No. of tone-one monosyllabic words: 1096
No. of tone-two monosyllabic words: 1132
No. of tone-three monosyllabic words: 995
No. of tone-four monosyllabic words: 646
No. of tone-five monosyllabic words: 1196
No. of tone-six monosyllabic words: 1028
No. of tone-seven monosyllabic words: 1163
No. of tone-eight monosyllabic words: 1004

Using CISS-Zhuang, the authors are able to access the individual syllables of a word. This capability is useful for data manipulation and collation involving poly-syllabic words.

II. Spelling System of Zhuang Language

II.1 The Alphabet

The writing system of the Zhuang language uses the letters of the Roman alphabet. But, as Zhuang is a tonal language with eight different lexical tones, some letters, all but one of them used solely for this purpose, are placed at the end of the syllables to indicate the tones they carry. Twenty-six letters (cf. the English alphabet) are used. The vowel letters are "a", "e", "i", "o", "u" and "w'"(!). The tone letters are "j", "q", "x", "z" and "h". "h" is also used as a consonant letter. All other letters are consonant letters.[5]

The syllable structure of Zhuang can be described adequately following the time-honoured Chinese tradition of analysing syllables in Sino-Tibetan languages into initials and rhymes.

INITIALS

b	mb	m	f	v
[p]	[b]	[m]	[f]	[v]

d	nd	n	s	l
[t]	[d]	[n]	[s]	[l]

g	gv	ng	h	r
[k]	[kv]	[ŋ]	[h]	[ɣ]

c	y	ny	ngv	by	gy	my
[ɕ]	[j]	[nj]	[ŋv]	[pj]	[kj]	[mj]

RHYMES

a	e	i	o	u	w
[a:]	[e:]	[i:]	[o:]	[u:]	[ɯ:]

ai	ae	ei	oi	ui	wi
[a:i]	[ai]	[ei]	[o:i]	[u:i]	[ɯ:i]

au	aeu	eu	iu	ou	aw
[a:u]	[au]	[e:u]	[i:u]	[ou]	[aɯ]

am	aem	em	iem	im	om	oem	uem	um
[a:m]	[am]	[e:m]	[i:m]	[im]	[o:m]	[om]	[u:m]	[um]

an	aen	en	ien	in	on	oen	uen	un	wen	wn
[a:n]	[an]	[e:n]	[i:n]	[in]	[o:n]	[on]	[u:n]	[un]	[ɯ:n]	[ɯn]

ang	aeng	eng	ieng	ing	ong	oeng	ueng	ung	wng
[a:ŋ]	[aŋ]	[e:ŋ]	[i:ŋ]	[iŋ]	[o:ŋ]	[oŋ]	[u:ŋ]	[uŋ]	[ɯŋ]

ap	aep	ep	iep	ip	op	oep	uep	up
ab	aeb	eb	ieb	ib	ob	oeb	ueb	ub
[a:p]	[ap]	[e:p]	[i:p]	[ip]	[o:p]	[op]	[u:p]	[up]

at	aet	et	iet	it	ot	oet	uet	ut	wet	wt
ad	aed	ed	ied	id	od	oed	ued	ud	wed	wd
[a:t]	[at]	[e:t]	[i:t]	[it]	[o:t]	[ot]	[u:t]	[ut]	[ɯ:t]	[ɯt]

ak	aek	ek	iek	ik	ok	oek	uek	uk	wk
ag	aeg	eg	ieg	ig	og	oeg	ueg	ug	wg
[a:k]	[ak]	[e:k]	[i:k]	[ik]	[o:k]	[ok]	[u:k]	[uk]	[ɯk]

Note: Very "broad" transcriptions are used. Short "a", "i", "o" and "w" are considerably more central than the corresponding long vowels.

TONES

1. The letters "z", "j", "x", "q" and "h" are used to represent tones. They are written at the end of syllables:

z - tone 2 (31), a low falling contour;[6]
j - tone 3 (55), a high level contour;
x - tone 4 (42), a high falling contour;
q - tone 5 (35), a high rising contour;
h - tone 6 (33), a mid level contour;

2. Tone 1 (24), with a low rising contour, is not marked.

3. Tone 7 (55 for a short nucleus and 35 for a long nucleus) and tone 8 (33), corresponding to upper and lower entering tones in Chinese, are represented by the letters "p", "t" and "k", and "b", "d" and "g" respectively.

II.2 Combination Constraints

Implicit in the account given above of the Zhuang alphabet are certain constraints for legal combination of letters in Zhuang syllables. First, certain sounds are represented by more than one letters. The voiced initial plosives, for example, are represented by "mb" and "nd"; "ie", "oe", "ae" (in certain contexts), "ue" and "we" are simple vowels; and "ng" also represents one sound. Second, under the heading 'RHYMES' are all the rhymes allowed. This tells us what combinations of vowels are possible in diphthongs, as well as what vowel combinations are go with what consonant codas. Third, the initials include onset-glide combinations like "gv". This also reflects phonotactic constraints.

Phonotactic constraints, which have not been given above, governing what initials, can legally be followed by what rhymes have not been considered. As CISS-Zhuang is used only for identifying legitimate syllables, this poses no serious problems, because, assuming that words are spelled correctly, letter combinations violating these constraints will not occur. However, in the case of such illegal combinations being presented to the system as a result, say, of spelling errors, the system will not be able to identify them as such. Another problem with this deficiency is that it will be impossible to use the system as a syllable generator. If it were so used, over-generation will result. When reliable information about these missing phonotactic constraints are available, the authors will consider

adding them to the system.

II.3 Demarcation Rules

Syllables in many words can be identified as there is only one solution that yields legal syllables as given in II.1. For example, coit (first day of the month) can only be co-it because oit is not a legal rhyme. Besides, syllable demarcation in the Zhuang spelling system is governed by the following rules:[7]

(1) If a consonant letter is both preceded and followed by vowel letters, then it goes with the following vowel letter by default. If the consonant letter is separated from the following vowel letter by the demarcation symbol "'" (the apostrophe), then it goes with the preceding vowel letter. For example, byagaq (a kind of fish) is bya-gaq and sim'in (a kind of feeling) is sim-in.

(2) If consonant letters occur one after another, and if they are preceded and followed by vowel letters, then, by default, the first consonant letter goes with the preceding vowel letter, and the second consonant letter goes with the following vowel letter. For example, banhaet (in the morning) is ban-haet. For both consonant letters to go with either the preceding vowel letter or the following one, the demarcation symbol "'" has to be used. For example, seng'eiq (business) and bi'gvaq (last year).

(3) The tone-marking letters "z", "j", "x" and "q" and the plosive coda letters "p", "t", "k", "b", "d", "g" serve to mark syllable boundaries. For example, biengz-beih (dragonfly) is biengz-beih and bakmbeau (talkative) is bak-mbaeu.

(4) When two or more vowel letters occur in succession, and if there is more than one way to put the syllable boundary, then the demarcation sign "'" will be used. For example, go'ien (tobacco). Similarly, "'" is also used to mark the syllable boundary in the case of 'ambiguous' sequences of three or more consonant letters. For example, cin'gya (relative by marriage).

III. Implementation of CISS-Zhuang system

The CISS-Zhuang system was implemented using the dBASE III PLUS programming language on a PC/AT compatible computer. Procedural programming techniques were used to implement a system following the augmented

transit network (ATN) approach.[8] The system will generally yield 'unabiguous' results in the sense that definite syllable boundaries are correctly identified. There are occasions where the network system will turn out ambiguous results. In such cases, disambiguation is carried by firing rules derived from the syllable demarcation rules given in II.3. The system also has the ability to identify some spelling errors as such, although this ability is not comprehensive in that it is possible for the system to accept certain illegal syllables (cf. II.2)

IV. Formalization of the Zhuang Spelling System

The approach adopted in the implementation of CISS-Zhuang is procedural, as an ATN necessarily is. However, the authors appreciate the advantages of a declarative approach. One of these advantages is that rules used in a declarative system are maintained as an separate and independent module of the system. The ensuing discussion is thus in terms of 'grammars',[9] though CISS-Zhuang has actually not be implemented as such.

IV.1 Context-Free Re-write Rules

We can conceive of a Context Free Grammar (CFG) as a formal representation for the Zhuang spelling system as described in II.1. In this grammar, a word is defined to be a linear concatenation of a number of syllables, normally not more then 6 in number, and a syllable may have four components, namely the Onset, Vowel, Coda[10] and Tone, represented by O, V, C and T respectively. There may be a delimiter "'" between two syllables, if the boundary cannot be determined without a such a mark. The initial symbol is W, standing for a word. Brackets are used to indicate optional elements in a rule, and numbers indicate positions within O, V or C. Lowercases and a single quotation mark in braces are all terminal symbols (Upper case letters are not discussed here, though they are properly dealt with in CISS-Zhuang). The whole grammar is constructed as follows.

```
W -> S (D) (W)
D -> {'}
S -> (O) V (C) (T)
O -> O1 (O2) (O3)
O1 -> {b,c,d,f,g,h,k,l,m,n,p,r,s,t,v,y}
O2 -> {b,d,v,g,y}
O3 -> {v}
```

```
V -> V1 (V2) (V3)
V1 -> {a,e,i,o,u,w}
V2 -> {i,u,w,e}
V3 -> {u}
C -> C1 (C2)
C1 -> {m,n,p,t,k,b,d,g}
C2 -> {g}
T -> {z,j,x,q,h}
```

The grammar covers all syllables in Zhuang language. But it also recognizes many concatenations of letters not considered in Zhuang as legal syllables. For example, "vvv" is regarded as a permissible onset as O1, O2 and O3 can each be a "v", though it is in fact not allowed in the language. Though our grammar is not intended to have the capability to reject all illegal spellings (cf. II.2), a sequence as blatantly illegal as "vvv" must be rejected.

IV.2 Constraints

In order to prevent the grammar from being too ready to accept illegal combinations, we can improve it by introducing constraints. These constraints, indicated by a "|", act upon the invocation of rules at consonant and vowel level. A constraint rule defines a legitimate context for a letter to show up. The constraints that need to be added to the grammar are as follows.

```
O2 -> {b|O1=m, d|O1=n, v|O1=g, g|O1=n, y|O1ϵ{n,b,g,m}}
O3 -> {v|O2=g}
V2 -> {i|V1<>i, u|V1ϵ{a,e,i,o}, w|V1=a, e|V1<>ė}
V3 -> {u|[V1=a & V2=e]}
C -> C1 (C2)|[V3=ϕ & V2ϵ{e,ϕ}]
C1 -> {m,p,b|[V3=ϕ & V2ϵ{e,ϕ} &V1<>w]}
C1 -> {n,t,d|[V3=ϕ & V2ϵ{e,ϕ}]}
C1 -> {k,g|[V3=ϕ & not[V1=w & V2=e]]}
C2 -> {g|[C1=n & not[V1=w & V2=e]]}
T -> {[z,j,x,q,h]|C1ϵ{m,n,ϕ}}
```

In the above formulation, we have added the terminal symbol "ϕ" to the rewrite rules given in IV.1. A constraint may contain more than one constraint equations, in which case they are put in square brackets. If any equation fails, the whole constraint will not be satisfied. A constraint of this kind is, in principle, much like a test in ATN to be checked in the process of traversing an input string. So, this constraint-based grammar is equivalent, in terms of its power, to an ATN.

IV.3 Problems with the Constraint-Based Grammar

The grammar described above should be able to recognize all legitimate syllables in Zhuang. It also has the ability to reject most illegal forms. We have already seen in II.2 that illegal syllables violating phonotactic constraints governing the co-currence of initials and rhymes will be considered legal in the system. Besides this, the grammar described above also fail to realize that "oe", "ue", "ie" and "we" only occur in closed syllables. That is, they must be followed by consonant letters in the syllable. CISS-Zhuang, as implemented by the authors, takes care of this by looking ahead to confirm the availability of an allowed coda letter. This will also lead us to the next section.

V. Look-Ahead Parsing in the Implementation of CCIS-Zhuang

V.1 Process of Recognition

The complete process of recognizing one syllable in CISS-Zhuang is divided into three steps, corresponding to the three stages of a syllable, namely:

(1) Recognition of the Head, consisting of not more than three onset consonant letters;
(2) Recognition of the Middle, consisting of not more than three vowel letters;
(3) Recognition of the Tail, consisting of coda consonant letters and tone mark letters.

Using the rules corresponding to the constraint-based grammar described in IV.1 and IV.2 (implemented in a procedural manner), the first two stages yield unamibguous results. A string is either legal or illegal depending on whether or not it can be parsed. The maximum match principle used in identification (or segmentation) of words in Chinese texts [11] is used in conjunction with the parsing rules. The maximum match principle resolves all ambiguities encountered in these two steps. For example, the string "aeu" be one vowel:"aeu", two vowels: "ae|u" or "a|eu", or into three: "a|e|u", but under the maximum match principle it will be taken as one vowel. The principle proves to conform to norms of the spelling scheme.

V.2 Ambiguities: Indeterminate Syllable Boundaries

In contrast to the first two stages, the third stage presents difficulties. Indeed, all ambiguities arise in this stage and they are all concerned with the boundary between two syllables. In determing a syllable boundary, the grammar (more correctly the rules) relies heavily on the use of the demarcation sign "'". Whenever this sign is there to tell the syllable boundary, the matter is decided. But, for example, the indeterminate syllable boundary in byagaq mentioned in II.3 will have to resolved in the absence of "'". Additonal rules on determination of boundaries need to be formulated in such circumstances.

V.3 Rules in Addition to Constraints

In order to resolve ambiguities concerning syllable boundaries, we build up a small knowledge base containing rules to supplement the "grammar" in CISS-Zhuang. We adopt a wait-and-see strategy and invoke look-ahead[12] rules to carry out disambiguation concerning the Tail part of a syllable. These rules derived from the demarcation rules given in II.3, apply when "'" is not available to resolve the ambiguity:

(1) If W(...VαV...), S(...V), S(...Vα), S(αV...), and not(αϵV), then choose W(S(...V)|S(αV...));

(2) If W(...VαßV...), S(...Vα), S(ßV...), and not(αϵV or ßϵV), then choose W(S(...Vα)|S(ßV...).

All remaining ambiguities will be declared errors by the system.

In the implementation of CISS-Zhuang, these rules, which are integrated with "grammar rules" and will be called in the look-ahead operation, constitute a small knowledge base. The rules necessary for the operations mentioned in IV.3 are also of a look-ahead nature.

VI. Conclusion

Making use of concepts and techniques used in natural language processing, the authors have thus developed a system that can identify individual syllables in words in the Zhuang language spelled in the standard romanized writing system. With this tool in their hands, they can now computerize research findings on the Zhuang lexicon and then collate them with each other with the confidence that they can access the individual syllables. As nearly all Zhuang syllables are themselves morphemes, this also means that the

system is a useful tool to access the individual Zhuang morphemes. The system has been implemented using a procedural approach. And, it should possible to develop a more flexible system using a declarative grammar approach.

V. Acknowledgement

The Zhuang lexicon project, for which CISS-Zhuang has been developed, is supported by Strategic Resarch Grant No. 7024 of the City Polytechnic of Hong Kong.

Notes

1. As at 1982. See Zhou (1990:17).
2. See, for example, Yuan (1960:ch. 9).
3. Lai and Cheung (1990). The dictionary used is Saw-loih Cuengh Gun (1984).
4. Lai and Cheung (1990).
5. The ensuing account of the Zhuang writing system, derived from Sawloih Cueng Guns (1984), is reproduced, with adaptations, from Lai and Cheung (1990).
6. Couplets like 31 for tone 1 follows a five-point scale to indicate pitch and tone contour, with 5 being the highest pitch and 1 being the lowest. 31 means a contour from mid to low, hence low falling.
7. Derived from Sawloih Cuengh Gun (1984).
8. See Woods (1987).
9. See Gazdar & Mellish (1989: ch. 4), Partee et al. (1990: 433-5-6).
10. See Hyman (1975: 188 ff).
11. Liang (1987), Kit et al. (1989).
12. Marcus (1980).

References

Gazdar, G. and C. Mellish (1989) Natural Language Processing in Prolog. Wokingham: Addison Wesley.

Hyman, L.M. (1975) Phonology: Theory and Analysis. New York: Holt, Rinehart & Winston.

Kit, C., Liu Y. and N. Liang (1989) "Automatic Segmentation of Chinese Texts". Chinese Information Science Journal. Beijing 1990 Vol 2, 1-8.

Lai, B.Y. and C.W. Cheung (1990) "A Study of Cognates and Loanwords in the Zhuang Language and Cantonese Chinese with the Help of a Computerized Database". Paper presented at International Conference on Theoretical and Applied Studies of Chinese and English, Hong

Kong, June 1990.

Lai, B.Y., Lun, S., Sun, C.F. and M. Sun (1991) "A Maximal Match Segmentation Algorithm Using Mainly Tags for Resolution of Ambiguities for Chinese Textx". Proceedings of ROC Computatinal Linguistics Conference. Taipei, 135-146.

Liang, N. (1987) "Chinese Text Segmentation System - CDWS". Chinese Information Science Journal. Beijing 1987 Vol 2,44-52.

Marcus, M.P. (1980) A Theory of Syntactic Recognition for Natural Language. MIT Press:Cambridge, MA.

Partee, B.H., ter Meulen, A. & R. E. Wall (1990) Mathematical Methods in Linguistics. Dordrecht: Kluwer Academic Publishers.

Sawloih Cuengh Gun (A Zhuang-Chinese Dictionary).Guangxi: Guangxi Minority Nationalities Press, 1983.

Winograd, T. (1983) Language as a Cognitive Process. Vol. I: Syntax. Wesley:Reading, MA.

Woods, W.A. (1970) "Transition network grammars for natural language analysis". Communication of ACM, 13, 591-6. Reprinted in Grosz, B.J., Jones, K.S. & B,L. Webber (eds.) (1986), Readings in Natural Language Processing. Los Altos: Morgan Kaufmann, 71-87.

Woods, W.A. (1987). "Augmented Transition Network Grammar". In Shapiro, S.C. (ed.), Encyclopedia of AI . New York:Wiley, 323-33.

Yuan, J. (1960) Hanyu Fangyan Gaiyao (The Chinese Dialects). Peking: Writing Reform Press.

Zhou, Y. (1990) "Reflections on 'Unified Writing System for Zhuang and Puyi'". Minority Languages. Beijing 1990 Vol 2, 16-26.