Some puzzles in Cao Lan

Kenneth J. GREGERSON & Jerold A. EDMONDSON
University of Texas at Arlington

There is a Latin phrase tertium quid for something that escapes easy division into two groups supposed to be exhaustive.¹ In this paper we outline why the Cao Lan language may be regarded as a tertium quid in regard to:

its tonality--Cao Lan appears to have merged its tones so that today as few as three may remain in live syllables and two in dead syllables, but this radical surgery has in part been compensated for by developing vowel quality and voice quality contrasts,

its phonology--Cao Lan demonstrates some phonological rules regarded as exclusively Northern Tai (NT) and some others that have been considered as Central Tai (CT) features,

its vocabulary--Cao Lan has a sizeable inventory of vocabulary that is common Tai, but also it has both Central Tai and Northern Tai forms as well.

The motivation for these variegate features we believe lies in the migratory history of the Cao Lan people. Cao Lan by their own account and from the study of experts is a group that has moved from its original homeland in NT territory in China to CT territory in Vietnam. In this paper we will present some evidence from our own fieldwork and computer-assisted analysis subsequent to it about the frailty of Cao Lan tone as well as the mottled phonological and lexical traits in Cao Lan and speculate about their origins. First of all though, we present some background about this group.

1. Introduction

The Cao Lan-San Chây people are one of the fifty-four officially recognized minority nationalities of Vietnam. The hyphenated designation as a part of their name is a recognition that this group speaks two languages and has two non-overlapping cultures. The Cao Lan, for one, use a language that belongs to the Tai Branch, whereas the San Chây make use of a form of Han Chinese. As a whole, the

¹The research reported on here was supported by a grant to the authors from the National Endowment for the Humanities and the National Science Foundation for a project called “Languages of the Vietnam-China Borderlands, 95-97.”
Cao Lan-San Chây are reported to have a population of 114,000 according to the 1990 census, living concentrated in Tuyên Quang and Bắc Thái Provinces. They are also found in smaller numbers in: Yên Bái, Thái Nguyên, Vĩnh Phúc, Bắc Giang, Lạng Sơn, and Quảng Ninh Provinces. Our Cao Lan informant, Mr. Hoàng Trưởng Vinh, used for his people the autonym San Chây, which surprised us in as much as San Chây is the official name not of his but of the Han-speaking group. From him we also learned that there is the unusual situation of two speech communities within one ethnicity, both using Chinese characters to record their folk customs and practices. And, as Mr. Vinh confirmed to us, at least some of the Cao Lan call themselves San Chây. MVNP 1978 reports that the Cao Lan and San Chây do not live in a classical diglossic situation of high language vs. low language, but as two groups with mostly different identities despite a small overlap today and a common link in the past. Looking at this history, scholars have concluded that at one time these people lived in China in territory along the border areas of Hunan, Guangdong, and Guangxi Provinces. They came to Vietnam circuitously arriving about 400 years ago and the unusual language use feature of the Cao Lan-San Chây appears to have resulted from events that occurred during this migration.

As we noted, the San Chây today speak mostly in a kind of Chinese, but some of their old people can also speak a Tai language. Similarly, seniors among the Cao Lan, such as Mr. Vinh, can speak and write in the Han language. Despite the language choice difference, the two have nevertheless common features of culture, language, and traditions. Some octogenarians among the Cao Lan suggest that they were originally San Chây; and in parallel fashion some older members of the San Chây community suggest they were once Cao Lan. They do share common folkloric traditions and common writing, as is portrayed in one song called “Sinh Ca” (Ca = song). According to the people themselves, they arrived in Vietnam as one group. The late French scholar, André Haudricourt 1973, explained that after they began their migrations, they stopped in Guangxi for a time and assumed a local form of the Zhuang (Tai) language. Later, on the way to Vietnam, some of them adopted the written language of Chinese and some began to use a spoken form of Chinese. To add further to the complication, it has also been suggested that the Cao Lan-San Chây have a connection to the Yao people as well. Indeed, there is a group of Yao in Fangcheng in Guangxi who call themselves [san33 tcai33], Mao et al (1982:8). All these facts tell us that the two were in some sense one nationality with two partially overlapping speech communities whose original bilingualism has developed into separated mostly monolingualism through separation, as the majority of the San Chây live in Quảng Ninh and the Cao Lan live mostly in Tuyên Quang, Thái Nguyên, and Bắc Giang.
2. History of previous work

Cao Lan was analyzed first by Bonifacy. Later references are then in Haudricourt (1960) The most famous among the unpublished sources is the 1938 survey of the École française d'extrême orient (EFEO) of 1938. It is a compilation of about 500 items transcribed in Quốc Ngữ script by civil servants, members of the institute, and others gathered at various locations. Among these materials is the entry XVI.2 Màn Cao Lan of Yên Sơn Tuyên Quang, XI.7 Màn Cao Lan Phù Thọ, and XI. Màn Cao Lan Tieu Á. The first location is the same as the one being reported on here, moreover, the comparison below shows that the varieties of Cao Lan are the same.

Most of the post 1945 work has concentrated on the question of genetic affiliation, specially whether Cao Lan is a Northern Tai or a Central Tai language. Haudricourt 1960 suggested that Cao Lan was neither (Black or White) Thái nor Yao, and, moreover, to use the terminology of Li 1977 and Gedney it is neither Central Tai nor is it Northern Tai. Basically, it has phonological and lexical features of both. As far we know, there have not been other studies of the linguistic features of the language and very little information about it has found its way to print. In our modest paper on the Cao Lan language we will provide considerably more data than has heretofore been available on Cao Lan, examine some of its features with computer-assisted techniques, and highlight the features of the mixture. Our informant was Mr. Hoàng Trưởng Vinh of Tuyên Quang Province, Yên Sơn District, Kim Phú Village, and Giếng Tanh Hamlet. Mr. Vinh was born in 1926 and was 79 years of age at the time of the interview in 1995. He was a very good speaker of the Cao Lan language, using it on a daily basis with his family. He was able to read and write Chinese characters to render his language in written form.

We conclude from the study of Mr. Vinh's speech, very much in the sense of Haudricourt, that the special properties of Cao Lan suggest a profound and prolonged interfusion of Northern Tai and Central Tai influences with a significant adstratum of a kind of Chinese, presumably Pinghua, the first kind of Han spoken in Guangxi, cf. below. We did not find evidence of Yao influence in the language per se. The presence of Pinghua we believe to be crucially responsible for the special tonal and voice quality features of Cao Lao. In fact, we see many commonalities between Cao Lan and E, a mixed language reported on in Edmondson 1992.

3. Tonal system

The Cao Lan language has a very unusual prosodic system for a Tai language. In particular, we found that contemporary form showed a collapse of
many original categories into only a few tones. Indeed, the pitch alone seems to distinguish only three categories today, cf. below for the CECIL plots of the tones. Consider the following data:

A1: $tu^{31}$ mo$^{42}$ ‘pig’; $phon^{42}$ ‘rain’; $tu^{31}$ ma$^{42}$ ‘dog’; na$^{42}$ ‘thick’
A2: $tu^{31}$ pa$^{42}$ ‘fish’; $tu^{31}$ ka$^{42}$ ‘crow’; ma$^{42}$ ‘thread’; $tha^{42}$ ‘eye’;
thai$^{42}$ ‘to die’; ka:ri$^{42}$ ‘far’
A3: $?bon^{42}$ ‘sky, heavens’; $?dai^{42}$ ‘good’; $?ben^{42}$ ‘lightweight’;
$?jut^{42}$ ‘medicine’
A4: $tu^{31}$ va:i$^{31}$ ‘buffalo’; ha$^{31}$ ‘cogongrass’; na$^{31}$ ‘wet field’
B1: ma$^{42}$ ‘new’; $tu^{31}$ thu$^{42}$ ‘rabbit’; pha$^{42}$ ma$^{31}$ ‘palm hand’
B2: kai$^{42}$ ‘chicken’; $tha^{42}$ ‘shuttle of loom’
B3: $?ba^{42}$ ‘shoulder’; $?ca:n^{42}$ ‘soft’; za$^{42}$ ‘body’; $?da^{42}$ ‘scald’
B4: ta$^{42}$ ‘river’; poh$^{42}$ $?ba:u^{42}$ ‘older brother’
C1: thai$^{35}$ ‘intestine’; heu$^{35}$ ‘teeth’; na$^{31}$ ‘face’
C2: kon$^{35}$ ‘hip’; tai$^{35}$ ‘below’; hai$^{35}$ ‘near’
C3: $?ba:n^{35}$ ‘village’; za$^{35}$ ‘winnowing basket’
C4: ma$^{42}$ ‘horse’; so$^{41}$ ‘early’; kau$^{42}$ ‘owl’
D1L: ma:k$^{45}$ ‘fruit’; tha:p$^{45}$ ‘carry on pole’
D2L: po$^{45}$ ‘lungs’
D3L: $?be:k^{45}$ ‘carry on shoulder’; zik$^{45}$ ‘to peel’
D4L: lo:k$^{34}$ ‘to skin’; no:k$^{34}$ ‘outside’; lok$^{44}$ ‘child’
D1S: phak$^{43}$ ‘vegetable’; lap$^{45}$ ‘to pull’; hap$^{45}$ ‘to bite’
D2S: tap$^{45}$ ‘liver’; $tu^{31}$ kop$^{45}$ ‘frog’
D4S: sak$^{44}$ ‘to wash (clothes)’; mot$^{44}$ ‘ant’

We have prepared plots using WINCECIL and a compositing program. In Figures 1-5 we show the pitch trajectories for representative vocabulary illustrating A1 vs. A4 in Figure 1; B3 vs. B4 in Figure 2; C2 vs. C4 in Figure 3, D2S, D4S, D3L vs. D4L in Figure 4, and finally in Figure 5 we compare A1, B4, and C4 tone examples, ma$^{42}$ ‘dog’; ta$^{42}$ ‘river’; and ma$^{42}$ ‘horse’.

---

2CECIL and WINCECIL are hardware and software speech analysis systems (JAARS International, Inc., Waxhaw, NC) that allow reliable extraction of pitch in a field setting.

The following list uses the system of Gedney to portray the potential sources for tones. Gedney assumes five original tones called A, B, C, DL, and DS and four consonant types for the initials, called 1=voiceless friction (aspiration and voiceless sonorants and fricatives, 2=plain voiceless stops, 3=“preglottalized” voiced stops and glottal stop initials, and 4=original fully voiced stops. We use then A1 to categorize vocabulary with voiceless friction initials in proto tone A, etc.