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

[1]The 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 Phú, Bắc Giang, Lạng Sơn, and Quảng Ninh Provinces. Our Cao Lan informant, Mr. Hoàng Trường 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 [san^{33} $tɕai^{33}$], 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 Phu Thọ, and XI. Màn Cao Lan Tieu Ẩ. 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 Trường 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:[2]

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'; *mai*^{42} 'thread'; *tha*^{42} 'eye'; *tha:i*^{42} 'to die'; *ka:i*^{42} 'far'
A3: *ʔbon*^{42} 'sky, heavens'; *ʔdəi*^{42} 'good'; *ʔbəu*^{42} 'lightweight'; *ʔjuɯ*^{42} 'medicine'
A4: *tu*^{31} *va:i*^{31} 'buffalo'; *ha*^{31} 'cogongrass'; *na*^{31} 'wet field'
B1: *mɔ*^{42} 'new'; *tu*^{31} *thu*^{42} 'rabbit'; *pha*^{42} *məi*^{31} 'palm hand'
B2: *kai*^{42} 'chicken'; *thau*^{42} 'shuttle of loom'
B3: *ʔba*^{42} 'shoulder'; *ʔɔ:n*^{42} 'soft'; *zaŋ*^{42} 'body'; *ʔda*^{42} 'scold'
B4: *ta*^{42} 'river'; *pəi*^{42} *ʔba:u*^{42} 'older brother'
C1: *θai*^{35} 'intestine'; *hɛu*^{35} 'teeth'; *na*^{35} 'face'
C2: *kon*^{35} 'hip'; *tai*^{35} 'below'; *hai*^{35} 'near'
C3: *ʔba:n*^{35} 'village'; *zaŋ*^{35} 'winnowing basket'
C4: *ma*^{42} 'horse'; *so*^{31} 'early'; *kau*^{42} 'owl'
D1L: *ma:k*^{45} 'fruit'; *tha:p*^{45} 'carry on pole'
D2L: *pɔt*^{45} 'lungs'
D3L: *ʔbe:k*^{45} 'carry on shoulder'; *zik*^{45} 'to peel'
D4L: *lɔ:k*^{34} 'to skin'; *nɔ:k*^{34} 'outside'; *lək*^{34} 'child'
D1S: *phak*^{45} 'vegetable'; *ləp*^{45} 'to pull'; *hap*^{45} 'to bite'
D2S: *tap*^{45} 'liver'; *tu*^{31} *kop*^{45} 'frog'
D3S: *ʔdip*^{45} 'alive'; *ʔuk*^{45} 'brain'
D4S: *sak*^{34} 'to wash (clothes)'; *mot*^{34} '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'.

---

[2]CECIL 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.

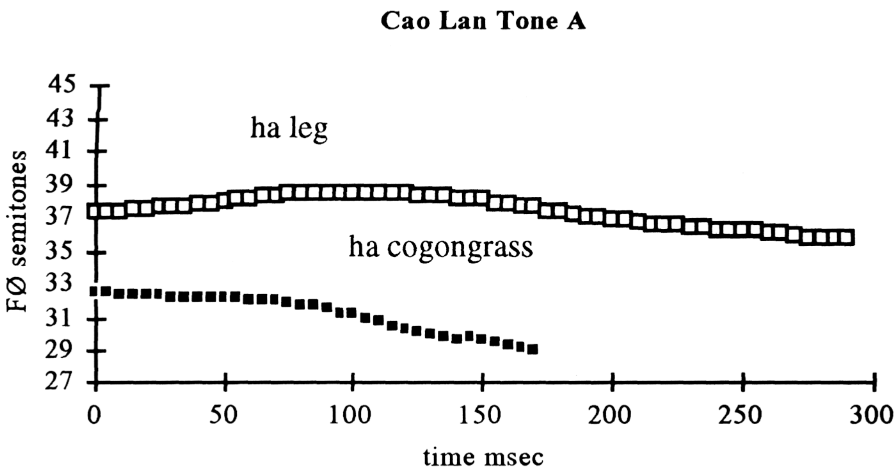


Figure 1. Cao Lan reflexes of Proto-Tai A tone

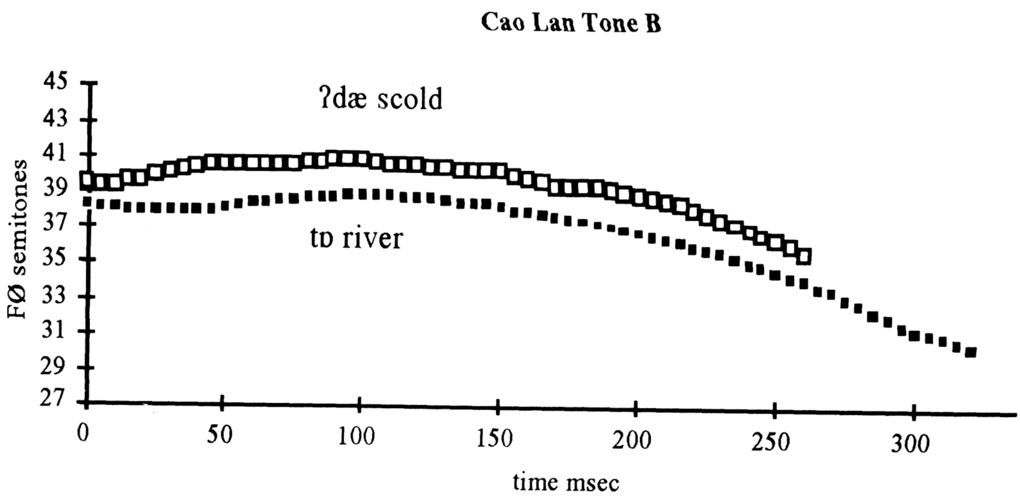


Figure 2. Cao Lan reflexes of Proto-Tai B tone

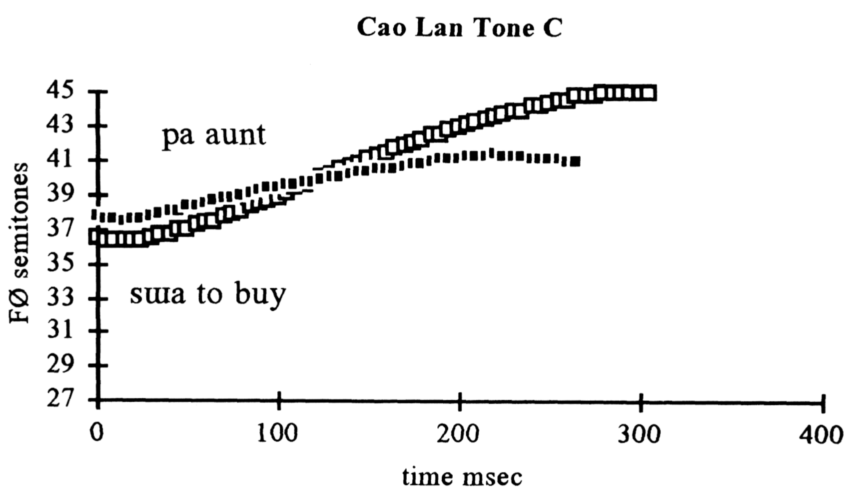


Figure 3. Cao Lan reflexes of Proto-Tai C tone

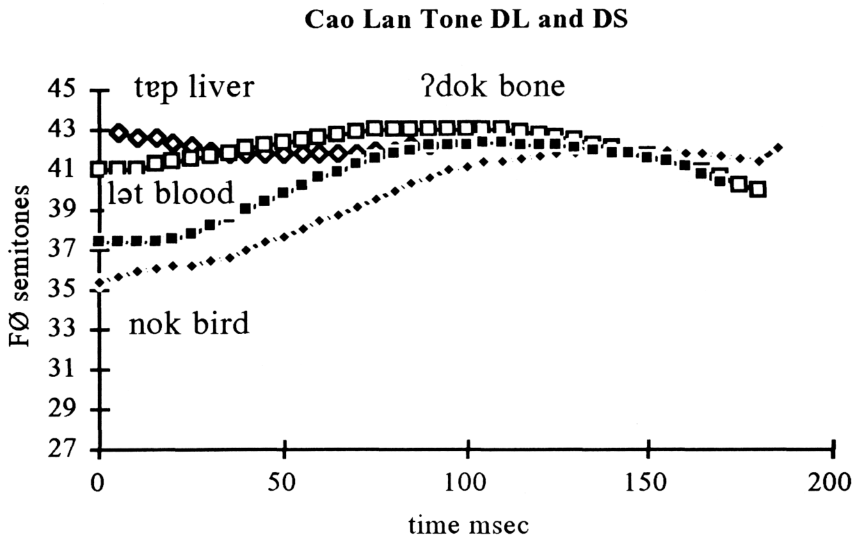


Figure 4. Cao Lan reflexes of Proto-Tai D tone

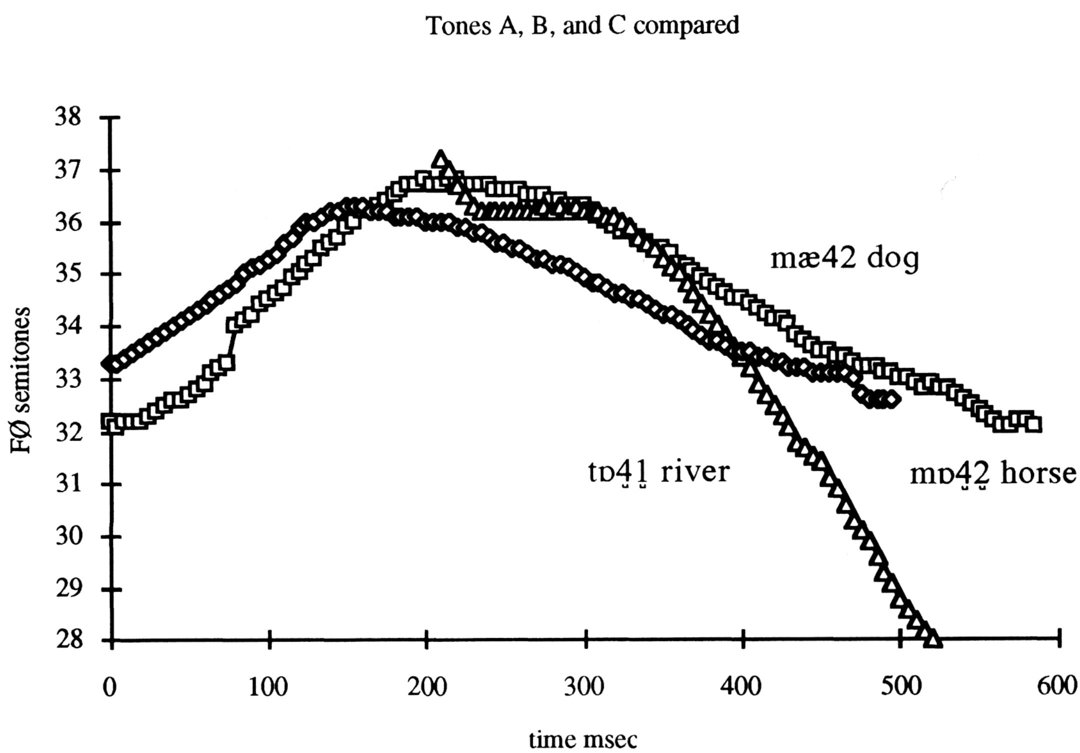
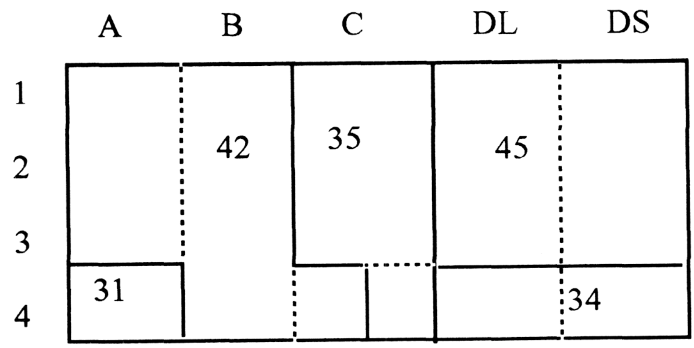


Figure 5. Comparison of *mæ* 'dog' and *mɒ* 'horse'

We can display the results of this comparison with a Gedney diagram showing the pattern of tone splitting, as is shown in Figure 6;



It is to be noted that some lexical items from C4 have become like C123, e. g. *məi*^{35} 'tree' whereas others resemble the pitch trajectories of the categories A123 and B1234. While the pitch may be similar, that is not to say that the words are homophonous. Indeed, Mr. Vinh repeatedly pointed our errors in pronouncing the items *ma*^{42} 'dog' vs. *ma*^{42} 'horse'. We finally realized that in his speech an expected tonal contrast would better be called one of voice quality and vowel quality transcribable as *mæ*^{42} 'dog' vs. *mɒ*^{42} 'horse'. The vowel in dog is higher and further front than that in horse. In below we display a plot of vowel formants determined by examining spectral data (F2-F1 vs. F1), which clearly demonstrates for several locations in each syllable the quality of the two vowels involved.

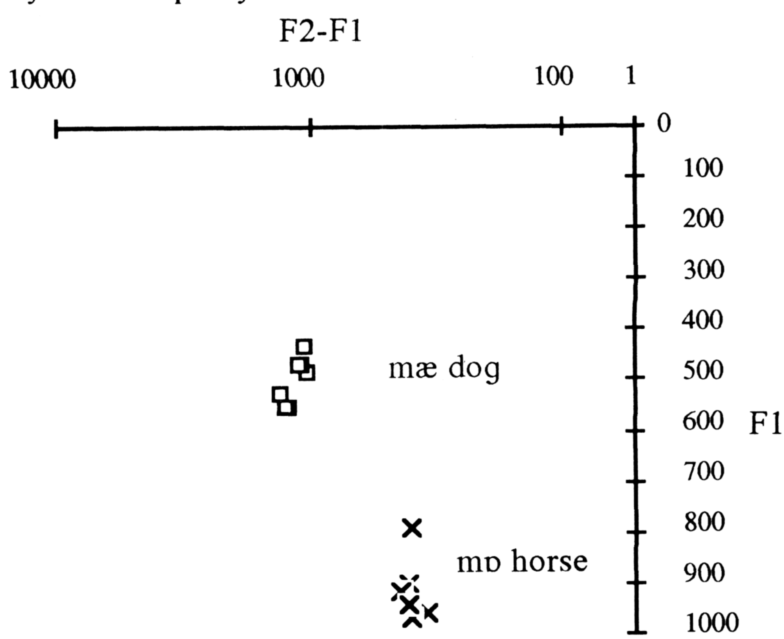


Figure 7. Comparison of the vowel formants of *mæ* 'dog' vs. *mɒ* 'horse'

Once alerted, we examined the voice quality (tense vs. lax) and vowel quality (high-front vs. low-back) to see if there is systematic use of these phonological contrasts to enhance or replace a weakened tonal system. We have concluded at this point that though there is a tendency for high tone items to have tense voice and higher/fronter vowels, there are also numerous exceptions to this tendency. It cannot be said that Cao Lan has developed a register system of the manner found in some Mon-Khmer languages.

4. The phonological and lexical puzzles

Beyond the strange tone features, the Cao Lan language has also developed unexpected segmental features in its phonology and vocabulary. Cao Lan—like Nùng An of Cao Bằng and Ts'ưn -wa of Quảng Ninh Province, as Strecker reports—possesses characteristics that are geographically out of place; they seem half like the languages of the Northern Tai (NT) Subbranch of Tai and half like those of the Central Tai (CT) Subbranch. Specifically, Strecker wants to argue that Cao Lan belongs to the Northern Tai Branch of Tai or Dioi, because in Cao Lan, like NT and unlike CT languages **dl* clusters, as in **dluon^{A} -> lom^{31}* 'wind' result in the same initial as **l* or **r*, as in **rɯan^{A} -> laan^{31}* 'house'. Strecker regards the development of **thr-* and **tr-* to *th-* (according to Li Fang Kuei's 1977 reconstruction), as in **tra^{A} -> tha^{42}* 'eye' and **thrau^{A} -> thau^{42}* 'louse' to be "areal" properties and not genetically inherited properties. However, from our data the picture of genetic affiliation of Cao Lan seems much murkier. Below we present fuller data than Strecker had at his disposal on the developments of the proto **r-* and **l-* as well as those of the proto clustered *r*: **dl-*, **thr-*, **nl-*, **kr-*, **khr-*, **gr-*, **pr-*, **pl-*, as well as **f-*, **fr-/vr-/vl-*, **hr-*, and **hw-*. We have provided in many cases corresponding data from a clear example of a NT language in Vietnam, namely, Giấy of Lào Cai Province and a clear example of a CT language of Vietnam, namely, Tày Cao Bằng.[3]

The development of **dl-*

Gloss	Cao Lan	Giấy Lào Cai (NT)	Tày Cao Bằng (CT)
'wind'	lom^{2}	ðum^{2}	lom^{2}
'fingernail'	lip^{8}	zit^{8} fɯŋ^{2}	lep^{8}

[3]Tones are marked following the Chinese system for recording tones in Tai languages as: A (high)=1; A (low)=2; B (high)=5; B (low)=6; C (high)=3; C (low)=4; DL (high)=9; DL (low)=10; DS (high)=7; and DS (low)=8.

The development of **dr-*

Gloss	Cao Lan	Giấy Lào Cai (NT)	Tày Cao Bằng (CT)
'boat'	lu²	ðua²	lɯa²
'day a. tomorrow'	ləi²	ðɯ²	lɯ²

The development of **r-*

'house'	laan²	ðaan²	rɯən²
'rice husk'	lam²	ðam²	ram²
'dry field'	lai⁶	ði⁶	rei²

The development of **hr-*

'to bark'	lau⁵	ðau⁵	hau⁵
'mushroom'	let⁷	ðat⁷	---

The development of **hw-*

'comb'	loi¹	ðoi¹	wi¹

The development of **nl/r-*

'water'	nom⁴	ðɐm⁴	nɐm⁴
'outside'	nɔ:k¹⁰	ðɔ:k¹⁰	nɔ:k¹⁰
'bird'	nok⁸	ðɔk⁸	nɔk⁸

The development of **thr-*

'carry on a pole'	thaap⁹	ðaap⁹	thaap⁹
'tail'	thuŋ¹	ðɯŋ¹	thaŋ¹
'stone'	thən¹	ðin¹	thin¹
'loom'	thok⁷		thuk⁷
'louse'	thau¹	ðau¹	thau¹
'to cook'	thoŋ¹	ðuŋ¹	thoŋ¹
'carry hanging'	thiu³	ðiu³	thiu³
'hailstone'	thet⁷	---	thep⁷

As one can see from this data, the NT language Giấy demonstrates the interdental fricative [ð-] as a reflex of all of these proto-initials, whereas Cao Lan has **dl-, *dr-, *r-, *hr-, *hw- -> l-*; but **nl/r- -> n-*; and **thr-* becomes *th-*. Finally, Tày Cao Bằng has the richest set of reflexes: **dl-* and **dr- -> l-*; **r-* remains **r-;* **nl/r-- -> n-*; and **thr-* becomes *th-*. So, it appears to us that Cao Lan occupies a middle ground between maximum merger--in the case of Giáy--and maximum preservation--in the case of Tày Cao Bằng. Or it seems about half way between a NT and a CT language and it would be somewhat arbitrary to see some developments as "areal" and some inherited.

In regard to aspiration, the loyalties of Cao Lan are again split. As Li 1977 has pointed out, NT languages seem to have lost aspiration in stops at an early time without engendering any changes in the tonal system. On this test too, Cao Lan shows itself to be neither fish nor fowl, aspirated or non-aspirated. Consider:

'near'	$khai^3$	$tɕaɯ^3$	$səɯ^3$
'vegetable'	$phak^7$	$pjik^7$	$phjak^7$

But the number of unaspirated or NT-like forms predominates in the lexicon so far studied:

'kill'	ka^3	ka^3	kha^3
'walk'	$pəəi^3$	$pjaai^3$	$thaai^3$

On the scorecard then, the behavior of aspiration in Cao Lan basically argues for a NT heritage. However, there are equally well rules of CT that seem much in evidence in Cao Lan as well. For example, there is a rule that **f- -> ph-*. This rule operates widely in CT and SW Tai, e.g. in Shan, but it is virtually the norm in the CT languages of Vietnam.

The development of **f -> ph-*

'cloud'	$phɯ^3$	$vɯ^3$	pha^4
'cotton'	$pha{:}i^5$	$va{:}i^5$	$pha{:}i^3$
'rain'	$phon^1$	pun^1	$phən^1$
'hand palm'	pha^5	---	---

Another feature of NT vis-à-vis CT is that a number of common items are found in the lower tone set, whereas CT and SWT have these in the high set. On this litmus Cao Lan sides with the North.

'excrement'	$həi^4$	$ʔe^4$	khi^3
'rice'	$hɐu^4$	hau^4	$khɐu^3$

There are also some differences of tone or segmental elements that seem to be restricted to individual words. First consider the items in which Cao Lan resembles the CT languages (as in Tày Cao Bằng in the third column):

'ladder'	*ʔdəi¹*	*lai¹*	*ʔduɐi¹*
'this'	nəi³	ni⁴	nəi³
'a fly'	meŋ² ʔiŋ²	nɛŋ²	mɛɛŋ² wɐn¹
'to sleep'	nɔ:n²	nin²	no:n²

The following items, on the other hand, show similarities between Cao Lan and the NT language Giấy.

'rope'	saap⁹	sak⁹	zɯək⁹
'come back'	ma¹	ma¹	ma²
'snake'	ŋɯ²	ŋɯ²	ŋu²
'widow'	mai⁵	---	mai³
'meat'	nɔ⁶	no⁶	nɯa⁴
'snow'	nai¹		mɯi¹

Finally there are many items, reported in Zhang/Wei 1997 in which NT languages and CT languages have a lexical difference. Consider first these items:

'cloth'	pha:i⁵	paŋ²	pha:i³
'sheep'	ʔbe³	ji:ŋ²	ʔbe³
'goose'	pə:n⁶	ha:n⁵	pu:n⁶
'slippery'	mɐ:k⁸	mja:k⁸	ma:k⁸
'yellow'	lɯ:ŋ¹	he:n³	lɯ:ŋ¹
'mud'	tom¹	na:m⁶	tom¹
'bowl'	an³ toi³	wa:n³	tui³

But then again there are the items:

'sun'	thaŋ¹ ŋin²	tɕan³³ wan²	tha¹ wɐn²
'clothing'	pəu⁶	pu⁶	ɬɯa³
'sky'	ʔbon¹	ʔbɯn¹	fa⁴
'horn'	kɔk⁷	kok⁷	kau¹
'yesterday'	ŋin² lɯ:n²	lɯ:n²	wa²
'tiger'	kok⁹	kuk⁹	ɬɯa¹
'flower'	va²	ʔdok⁹ va²	ʔbjok⁷

5. Conclusion

In this paper we have reported on a very strange-looking tonal system and a phonology and lexicon with—on balance—equal indications of a NT and CT heritage. Admittedly, our ignorance about matters Cao Lan is still profound. Nevertheless, it seems to us that answers to the puzzling features of this language are to be sought in the fusion of three languages that must have existed in a speech community that was multilingual for some time. This would be a situation similar to that Edmondson 1992 reported for the language E of Sanjiang and Luocheng Counties, Guangxi. In that language we also found that the organization of tone contrasts was one of the casualties when fusion creolization occurs. The new system that results shows discontinuities in development without total language breakdown as is seen in fission creoles of island nations. Moreover, E also has a similar sort of voice quality development, presumably imported from Tuguai Hua, a local kind of Pinghua Chinese spoken in Guangxi Province.[4] The resemblance of the voice and vowel quality in Cao Lan and E is uncanny. So, we would venture to speculate that Pinghua figures in the history of Cao Lan. We suppose, the Cao Lan people were once a NT people moving southward who stopped and interfused with a CT speaking group at a time early enough for some late CT rules still to operate. This contact must have been more direct than that which has occurred in most langugaes of the area, involving living together and intermarriage. That is why it seems to us procrustean to force Cao Lan onto either a NT or CT category. Cao Lan is what it has become, namely mottled and mixed at a very deep level. What's wrong with being a tertium quid anyway?

[4]Liang Min, in a paper read at the 30th ICSTLL in Beijing, suggested that *Pinghua* be recognized as a new type of the Han language. Pinghua is the kind of Chinese spoken by descendents of the garrison troups on roads and waterways in Guangxi beginning from Qin times. It is now not accorded the status of Yue, Kejia, Min, etc among Chinese dialectologists. Moreover, it is easily mistaken for Yue (Cantonese), because of its rich tonal system (8 or 9 tones) and codas (*-p, -t, -k* are still present). However, Pinghua, on close examination, is distinct from Cantonese and, moreover, antedates the arrival of Cantonese by many centuries as Liang Min showed. It is also the donor language for the loan words in the minority languages of Guangxi Province.

REFERENCES

Bonifacy, August. 1907. "Étude sur les Cao Lan." *T'oung Pao ou Archives concernant l'histoire, les langues, la geographie et l'ethnographie de l'asie orientale*. Serie ii. Vol. viii. Leiden: E. J Brille, 429-438.

Edmondson, Jerold A. 1992. "Fusion and diffusion in E, Guangxi Province, China." In T. Dutton et al eds., *The Language Game: Papers in Memory of Donal C. Laycock* (Pacific Linguistic Series, C-110). Department of Linguistics, Australian National University, Canberra, 131-140.

Haudricourt, André. 1960. "Note sur les dialectes de la région de Moncay." *BEFEO* 50:167-177.

Haudricourt, André. 1973. "Mấy diếu nhận xét về lỳ luận và thực tiễn nhân một chuyển di tham các dàn tộc thiểu sổ ở Táy Bắc Việt Bắc." *Ngôn Ngữ* 3.5. [Some remarks about theory and practice on the occasion of a trip to visit the ethnic groups of northwest and northeast Vietnam.]

Li Fang Kuei. 1977. *A Handbook of Comparative Tai.* Honolulu: The University Press of Hawaii.

MVNP, Vietnam Social Sciences Committee (ed.) 1978. *Các dân tộc ít người Việt Nam (các tỉnh phía bác).* [Minorities of Vietnam, the northern provinces.] Hanoi: Social Sciences Publishing House.

Strecker, David. 1985. "The classification of the Caolan languages." In S. Ratanakul et al eds., *Southeast Asian Linguistic Studies Presented to André-G. Haudricourt*. Bangkok: Mahidol University.

Zhang Yuansheng and Wei Xingyun. 1997. "Regional variants and vernaculars in Zhuang." In Edmondson and Solnit eds., *Comparative Kadai: The Tai Branch*. Dallas: Summer Institute of Linguistics.