Lund University, Dept. of Linguistics
Working Papers 33 (1988), 193-214

The East Asian Code Shift

Jan-Olof Svantesson

In his article from 1961, 'The code shift in Old Norse', Bengt Sigurd showed that the historical phonological processes of umlauting, syncope and breaking in Old Norse resulted in a change in the type of phonological code by which the language transferred the semantic contents of its morphemes and words. He suggested that the information content of a word could be quantified using methods from communication theory and showed that due to the changes mentioned, the amount of information conveyed by one phoneme increased in Old Norse, while the length of morphemes decreased.

In this article, I suggest a method for quantifying the distribution of information capacity over the segments of a syllable (or of other units), which takes phonotactic rules into consideration. This is necessary since phonotactic rules restrict the possible combinations of phonemes into syllables and words (cf. Sigurd 1965), and such restrictions decrease the amount of information conveyed by a phoneme by making its occurrence more predictable.

I will apply this method to languages in the East Asian area, showing that a code shift has taken place there, which consists of transferring information from the outer parts of a syllable to its vowel kernel.

Phonological processes with this effect have occurred in all the major language families of the area: Sino-Tibetan (including Chinese), Altaic, Austroasiatic and Kadai (Tai). There are two main types of such processes: those that increase the number of possible contrasts in the stressed vowel (e.g. by introducing new contrasts of length, nasality, tone etc.), and those that decrease the contrasting capabilities of other segments (e.g. by reducing consonant clusters or unstressed vowels, or by decreasing the number of consonants). In this article, I will not be concerned with these processes as such, but rather with their consequences for the phonetic coding of the languages.

INFORMATION CAPACITY

In information (or communication) theory, founded by Claude Shannon in 1948, the amount of information transmitted by a channel using a code with a finite number of symbols is measured by its entropy, which is defined as $-\sum p_i \log p_i$, where p_i is the probability of symbol number i (approximated by its relative frequency). If, as is usually done in information theory, logarithms to the base 2 are used, entropy is expressed in bits, i.e. number of binary choices.

It can be proved that the maximum amount of information per symbol that can be transferred by a noiseless information channel using a code consisting of n symbols is equal to log n. This corresponds to the case when all p_i are equal to $1/n$. The intuitive plausibility of this can be seen by considering codes with 2^n symbols, which have the maximal entropy log $(2^n) = n$ bits, corresponding to the fact that n binary choices suffice to specify 2^n symbols.

In this article, a language is regarded as an information channel which transmits messages by means of a code consisting of segmental phonemes (which may carry suprasegmental features). I will not be concerned with the probabilistic concept of entropy, but rather with the quantity log n, the maximal amount of information per symbol that a code with n symbols can transmit. This quantity can be called the information capacity of the code. Entropy represents the part of this capacity which is actually used by a channel, and the difference between these two quantities is its redundancy.

Information theory is treated e.g. in Shannon 1948, Ash 1965 and Billingsley 1965. Works with more emphasis on applications, e.g. to linguistics, are Meyer-Eppler 1959 and Jaglom and Jaglom 1973.

How the information distribution over the segments of a syllable can be determined is best shown by an example, so I will first do the computations in detail for Chinese.

CHINESE

The code shift that has taken place in Chinese is illustrated here by comparing Middle Chinese, the (reconstructed) language of the Tang dynasty (7th to 9th Centuries A.D.), with two modern Chinese dialects (or languages), Standard Chinese, based on the Peking dialect, and Shanghai, belonging to the Wu dialect group.

For Middle Chinese I have used the reconstruction by Chen 1976. There are several reconstructions of Middle Chinese, but the differences between them are not of a kind that would disturb the results arrived at here in any serious way. For Shanghai, I have used Sherard 1980, but my analysis is different from his. In Standard Chinese, I have analysed the initial consonants [tɕ], [tɕʰ] and [ɕ] as allophones of /ts/, /tsʰ/ and /s/ before the glide /j/. The 'apical' vowels [ɿ] and [ʅ], found only after sibilants, are treated as allophones of /i/. The syllables [ɕi], [tɕi] and [tɕʰi] are analysed as /sji/, /tsji/ and /tsʰji/, respectively. The syllabic consonants in Standard Chinese and Shanghai are disregarded since they fall outside the system by not combining with other segments and thus contribute very little to the total information capacity of the language.

Since the syllable is a basic unit in Chinese phonology and morphology (almost all morphemes are monosyllabic), it is a natural unit for determining the information distribution. A Chinese syllable can be analysed as having the structure $(C_i)(G)V(C_f)$, where C_i and C_f are consonants, G a glide and V the vowel kernel, the only obligatory element. The different phonemes that can occur in these positions in the three languages are shown in Table 1, and the occurring combinations of them are shown in Table 2, where I have followed the table in Chen 1976:132 for Middle Chinese, and that in Sherard 1980:39 for Shanghai.

When the information capacity of each position is computed, the possibilities of combination with the elements of the preceding position is considered, but dependencies which go beyond adjacent phonemes are ignored. In this way, the most important phonotactic restrictions are taken into account without making the method unnecessarily complicated. When important discontinuous dependencies occur, e.g. in vowel harmony languages, the method has to be modified to account for them (see Mongolian below).

First, I will illustrate the method by computing the information distribution over a Standard Chinese syllable. In Standard Chinese, the C_i position is one of 18 different initial consonants or Ø, so its information capacity is log 19 = 4.248 bits. If the medial glide position were completely independent of the initial consonant, its information capacity would be log 4 = 2 bits. In that case the total information capacity of the sequence C_iG would be 4.248 + 2 = 6.248 bits, which corresponds to $2^{6.248} = 76$ C_iG combinations. This would mean that all possible combinations of 19 initial C_i's (including Ø) and 4 G's (also including Ø) occur (19·4 = 76), but that is not the case. For instance, labial consonants do not combine with rounded glides, and postalveolar and velar consonants do not combine with front glides. There are only 51 actually occurring combinations (Table 2:1).

This table shows that there are 6 C's (*ts*, *tsʰ*, *s*, *n*, *l* and Ø) that can be followed by all 4 G's, 2 (*t*, *tʰ*) that can be followed by 3 G's, 10 (*p*, *pʰ*, *m*, *tʂ*, *tʂʰ*, *ʂ*, *ʐ*, *k*, *kʰ*, *x*) that can be followed by 2 G's and 1 (*f*) that can be followed by only 1 G. A weighted information capacity of a medial glide, which takes into account the different combination abilities can therefore be computed as log ((6·4 + 2·3 + 10·2 + 1·1)/19) = log (51/19) = 1.424.

Table 1. Segmental phonemes in Chinese.

Initial consonants:

Middle Chinese:	Standard Chinese:	Shanghai:
Ø	Ø	Ø
p t c k ʔ	p t k	p t k
b d ɟ g	pʰ tʰ kʰ	pʰ tʰ kʰ
pʰ tʰ cʰ kʰ	ts tʂ	ts tɕ
ts tʂ tɕ	tsʰ tʂʰ	tsʰ tɕʰ
dz dʐ dʑ	f s ʂ x	f s ɕ h
tsʰ tʂʰ tɕʰ	ʐ	m n ŋ
s ʂ ɕ x	m n	l
z ʐ ʑ ɣ	l	
m n ɲ ŋ		
l		

Medial glides:

Middle Chinese:	Standard Chinese:	Shanghai:
Ø	Ø	Ø
w ɥ j	w ɥ j	w j

Vowels:

Middle Chinese:	Standard Chinese:	Shanghai:
i u	i y u	i y u
e ə	ɤ	ɪ ʏ
a ɑ	a	e ɤ o
		ɛ ɔ
		a

Final consonants:

Middle Chinese:	Standard Chinese:	Shanghai:
Ø	Ø	Ø
p t c k	n ŋ	ʔ
m n ɲ ŋ	w j	ŋ
w j		

Similarly, the information capacity of the vowel is $\log((1\cdot5 + 1\cdot3 + 2\cdot2)/4) = \log 3 = 1.585$ bits (cf. Table 2:2), and that of the final consonant is $\log((2\cdot5 + 3\cdot1)/6) = \log 2.167 = 1.115$ bits (Table 2:3).

The average length of each segmental position (expressed in number of phonemes per position) can be computed by weighting in the same way. In the Standard Chinese example, the average length of the initial consonant position is $18/19 = 0.947$ phonemes, since 18 of the possible 19 choices have the length 1 phoneme, while the 19th (Ø) has length 0. The length of the glide position is $(6\cdot3/4 + 2\cdot2/3 + 10\cdot1/2 + 1\cdot0)/19 = 0.570$ phonemes (6 initial consonants can be followed by all three glides and Ø, 2 can be followed by two glides or Ø, 10 can be followed by one glide or Ø, and 1 can be followed only by Ø). The vowel is always present, so its length is 1, and the weighted length of the final consonant is $(2\cdot4/5 + 3\cdot0)/6 = 0.267$.

Table 2:1. Combinations of initial consonants and medial glides in Chinese.

Middle Chinese:

	Ø	j	w	ɥ
Ø		+		+
p	+	+	+	+
pʰ	+	+	+	+
b	+	+	+	+
m	+	+	+	+
t	+	+	+	
tʰ	+	+	+	
d	+		+	
n	+	+	+	+
l	+	+	+	+
ts	+	+	+	+
tsʰ	+	+	+	+
dz	+	+	+	+
s	+	+	+	+
z	+	+	+	+
tʂ	+	+	+	
tʂʰ	+	+	+	+
dʐ	+	+		+
ʂ	+	+	+	+
ʐ	+			
c	+	+	+	+
cʰ	+	+	+	+
ɟ	+	+	+	+
ɲ	+	+	+	+
tɕ	+	+	+	+
tɕʰ	+	+	+	+
dʑ	+	+		+
ɕ	+	+	+	+
ʑ	+	+	+	+
k	+	+	+	+
kʰ	+	+	+	+
g	+	+	+	+
ŋ	+	+	+	+
x	+	+	+	+
ɣ	+	+	+	+
ʔ	+	+	+	+

Standard Chinese:

	Ø	j	w	ɥ
Ø	+	+	+	+
p	+	+		
pʰ	+	+		
f	+			
m	+	+		
t	+	+	+	
tʰ	+	+	+	
n	+	+	+	+
l	+	+	+	+
ts	+	+	+	+
tsʰ	+	+	+	+
s	+	+	+	+
tʂ	+		+	
tʂʰ	+		+	
ʂ	+		+	
ʐ	+		+	
k	+		+	
kʰ	+		+	
x	+		+	

Shanghai:

	Ø	j	w
Ø	+	+	+
p	+	+	
pʰ	+	+	
f	+		
m	+	+	
t	+	+	
tʰ	+	+	
n	+	+	
l	+	+	
ts	+		
tsʰ	+		
s	+		
tɕ	+		
tɕʰ	+		
ɕ	+		
k	+		+
kʰ	+		+
ŋ	+		
h	+		+

Table 2:2. Combinations of medial glides and vowel kernels in Chinese.

Middle Chinese:

	ɑ	a	ə	e	i	u
Ø	+	+	+	+	+	+
j		+	+			+
w	+	+	+	+	+	
ɥ		+	+			

Standard Chinese:

	a	ɤ	i	u	y
Ø	+	+	+	+	+
j	+	+	+		
w	+	+			
ɥ	+	+			

Shanghai:

	a	e	ɛ	ɪ	i	ʏ	y	u	o	ɤ	ɔ
Ø	+	+	+	+	+	+	+	+	+	+	+
j	+		+	+					+	+	+
w	+	+	+			+				+	+

Table 2:3. Combinations of vowels and final consonants in Chinese.

Middle Chinese:

	Ø	p	t	c	k	m	n	ɲ	ŋ	j	w
ɑ	+	+	+		+	+	+		+	+	+
a	+	+	+	+	+	+	+	+	+	+	+
ə		+	+		+	+	+		+		+
e		+	+	+		+	+	+		+	+
i	+										
u	+				+				+		

Standard Chinese:

	Ø	n	ŋ	j	w
a	+	+	+	+	+
ɤ	+	+	+	+	+
i	+				
u	+				
y	+				

Shanghai:

	Ø	ŋ	ʔ
a	+	+	+
e	+		
ɛ	+		
ɪ	+	+	+
i	+		
ʏ	+		
y	+		
u	+		
o	+	+	+
ɤ	+	+	+
ɔ	+	+	

What remains is to distribute the information contained in the suprasegmental phonemes, in this case the tones. Since there are four tones (I will limit myself to stressed syllables, and disregard the fact that there are unstressed syllables without tones), their total information capacity is log 4 = 2 bits. How these two bits are divided over the segments depends on one's views about the phonetic or phonological domain of the tones. According to traditional Chinese phonology, the domain of the tone is the rhyme, i.e. the vowel kernel and the final consonant, and this seems to be reasonably compatible with what is known about the domain of phonetic realisation and perception of the tones. I have chosen to distribute the information contents of the tones so that it is proportional to the average length of each of the positions involved (in this case the vowel and the final consonant), and also proportional to the proportion of voiced phonemes in the respective position, since voiceless phonemes cannot carry the tone. In Standard Chinese (but not Middle Chinese, see below) all final consonants are voiced. Then the distribution of the information due to the tones over the segments can be computed in this way: suppose that a tone-carrying position of unit length (and having only voiced phonemes) has the tonal information contents a. Then a is obtained by solving the equation:

$$a + 0.267a = 2,$$

giving a = 1.579, and the 2 bits of information contained in the tones are distributed so that the vowel gets 1.579 bits and the final consonant 1.579·0.267 = 0.421 bits.

The information distribution of Standard Chinese is summarised in Table 3, and is also illustrated in Figure 1. In the figure, the information distribution

Table 3. Information distribution in Chinese syllables.

Middle Chinese:

Position	Length	Information segm.	tone	total	%
C_i	0.972			5.170	40
G	0.718			1.874	14
V	1	2.000	+ 0.819	= 2.819	22
C_f	0.744	2.700	+ 0.365	= 3.065	24
Total	3.434			12.928	

I/L = 3.765

Standard Chinese:

Position	Length	Information segm.	tone	total	%
C_i	0.947			4.248	41
G	0.570			1.424	14
V	1	1.585	+ 1.579	= 3.164	30
C_f	0.267	1.115	+ 0.421	= 1.536	15
Total	2.784			10.372	

I/L = 3.726

Shanghai:

Position	Length	Information segm.	phonat.	tone	total	%
C_i	0.947	4.248	+ 0.229		= 4.477	45
G	0.298	0.706	+ 0.108		= 0.814	8
V	1	2.939	+ 0.363	+ 0.363	= 3.665	37
C_f	0.288	0.862	+ 0.052	+ 0.052	= 0.966	10
Total	2.533				9.922	

I/L = 3.917

over a syllable is plotted as a bar graph with one bar for each position, the width of the bar proportional to the average length, and its area proportional to its information capacity. The height of a bar is thus proportional to the information capacity per unit length (bit per phoneme).

Similarly, the information capacity (I) and average length (L) of the different positions in a Middle Chinese syllable can be computed in the following way (cf. Tables 1 and 2):

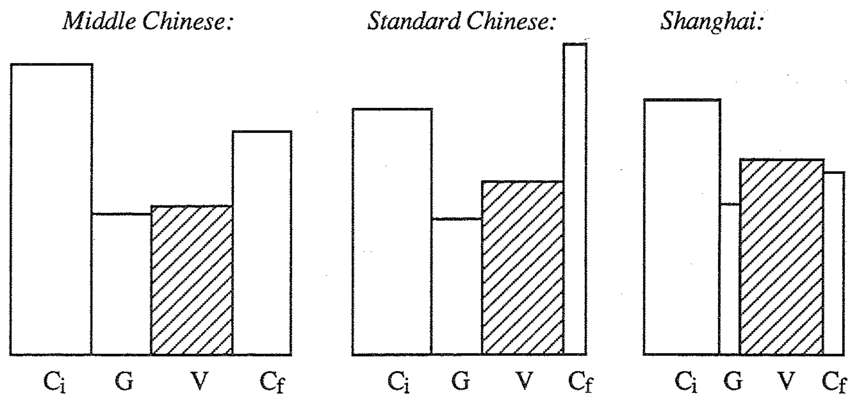

Figure 1. Information distribution in Chinese syllables.

Initial consonant: $I = \log 36 = 5.17$
$L = 35/36 = 0.972$

Medial glide: $I = \log((28\cdot 4 + 5\cdot 3 + 2\cdot 2 + 1\cdot 1)/36) = \log 3.667 = 1.874$
$L = (28\cdot 3/4 + 5\cdot 2/3 + 1\cdot 1/2 + 1\cdot 1 + 1\cdot 0)/36 = 0.718.$

Vowel: $I = \log((6 + 5 + 3 + 2)/4) = \log 4 = 2$
$L = 1.$

Final consonant: $I = \log((11 + 9 + 8 + 7 + 3 + 1)/6) = 2.700$
$L = 10/11 + 8/9 + 1 + 1 + 0 + 2/3) = 0.744.$

In Middle Chinese, there are three contrasting tones on open syllables and syllables ending in a voiced consonant, but no tonal contrast on syllables ending in a voiceless stop. Since there are 7 C_f's (including Ø) of the first kind and 4 of the second kind, the total information value of the tones is $\log((7\cdot 3 + 4\cdot 1)/11) = 1.184$ bits. According to the principle formulated above, if the vowel receives a bits of information from the tones, the final consonant should receive $0.744\cdot 4/10\cdot a$ bits (since its length is 0.744 phonemes, and 4 out of 10 final consonats are voiced). The quantity a is then obtained from the equation

$$a + 0.744\cdot 4/10\cdot a = 1.184,$$

giving $a = 0.819$.

The segmental information capacity of Shanghai is computed in the same way, and the result is shown in Table 3.

There are two suprasegmentals in Shanghai, phonation type and tone. If the initial consonant is an aspirate or /h/, there is no phonation type contrast, but syllables with other initials have an opposition between clear and breathy phonation. Thus, since 13 initials combine with two phonation types and 6 with one only, the information contents of the phonation type is $\log((13\cdot 2 + 6\cdot 1)/19) = 0.752$ bits. The phonation difference can be seen as a property of the whole syllable, since the initial consonants have different allophones depending on phonation type. Distributing the information contents of the phonation type over all segmental phonemes (weighting according to the principles given above) leads to the result shown in Table 3.

If phonation type is taken as the primary suprasegmental, a tone opposition is found only in syllables which have clear phonation and the final Ø or *-ŋ*, and such syllables have two tones, high or falling. In all other cases the tone is predictable from the combination of phonation type and final consonant:

	Final:	
	Ø, ŋ	ʔ
Phonation:		
Clear	high/falling	high
Breathy	rising	low

Thus the information content of the tone opposition is $\log((2\cdot 2 + 4\cdot 1)/6) = 0.415$. This is distributed over the vowel and final as shown in Table 3.

Comparison of these three Chinese languages shows that the modern dialects differ from Middle Chinese by having relatively more information capacity concentrated on the vowel kernel. While the Middle Chinese syllable has only 22% of its total information in the vowel, Standard Chinese has 30% and Shanghai 37%.

As shown below, this type of restructuring, which concentrates the information capacity on the vowel, has taken place in several East Asian languages, and for this reason I will refer to it as the East Asian code shift.

In Chinese, the code shift is the result of a number of well-known historical phonological changes, the most important being: (1) loss of some final consonants, accompanied by the introduction of new vowels, and (2) loss of distinctive voicing, leading to tone split, or to the development of phonemic phonation type.

These processes transferred information to the vowel by decreasing the number of contrasts in the consonants and, especially in Shanghai, by enlarging the vowel inventory.

It can be remarked that this redistribution of information within the syllable did not change the total information transfer efficiency (which may be measured by the amount of information per phoneme, I/L) appreciably, while the length of a syllable decreased, leading to a smaller amount of information being transferred by each syllable in Standard Chinese and Shanghai than in Middle Chinese. This is one possible explanation for the well-known fact that the number of disyllabic words is greater in these modern dialects than in Middle Chinese.

Although the earlier stages of Chinese are less well known than Middle Chinese, it is usually assumed that Old Chinese (the oldest records are from the 14th Century, B.C.) differs more from Middle Chinese than Middle Chinese does from the modern dialects. Old Chinese probably lacked tones and had initial and final consonant clusters, being phonotactically similar to its distant relative, Classical Tibetan (see below). Thus it can be assumed that a code shift involving transfer of information to the vowel was going on long before the Middle Chinese stage.

TIBETAN

Tibetan belongs to the Tibeto-Burmese branch of the Sino-Tibetan language family, which also includes Chinese. The earliest texts are from the 7th Century A.D., written with a phonemic alphabet, which is still used today, although the spoken language differs considerably from the written form.

Here I will compare modern Lhasa Tibetan with written, or Classical Tibetan, using the phonemic analyses and the comparison of Classical and Modern Lhasa Tibetan given in the introduction to the Tibetan-Chinese dictionary of Fù 1983.

There are many phonotactic and other phonological differences between Classical and Modern Tibetan: Classical Tibetan allows consonant clusters of up to four syllable initial consonants and two final consonants, while modern Tibetan has no consonant clusters. On the other hand, modern Tibetan has enlarged the simple five-vowel system of Classical Tibetan by adding more vowel qualities as well as contrasts of length and nasality, changes that were conditioned by the loss of final consonants. It also developed a four-tone system (Classical Tibetan has no tones). The segmental phonemes of the two languages are shown in Table 4.

Initial clusters in Classical Tibetan are shown in Table 5:1. In that table, '+' means that the cluster consisting of the consonant to the left followed by that shown on top exists, e.g. a 'j' means that the same cluster followed by *j* exists. It can for instance be seen that the clusters *bsk-*, *bskj-* and *bskr-* exist.

Final clusters consist of a non-dental consonant followed by *-s*, i.e. the existing clusters are *-bs*, *-ms*, *-gs* and *-ŋs*.

Table 4. Tibetan segmental phonemes.

Classical Tibetan

Initial consonants:

p	t	c	k	ʔ
p^h	t^h	c^h	k^h	
b	d	ɟ	g	
	s	ɕ		h
	z	ʑ		ɦ
	ts			
	ts^h			
	dz			
m	n	ɲ	ŋ	
	l			
	r			
w		j		

Vowels:

i		u
e		o
	a	

Final consonants:

Ø			
b	d		g
	s		
m	n		ŋ
	l		
	r		

Modern Lhasa Tibetan

Initial consonants:

Ø						
p	t			c	k	
p^h	t^h			c^h	k^h	
f	s	ʂ	ɕ	ç		h
	ts	tʂ	tɕ			
	ts^h	$tʂ^h$	$tɕ^h$			
m	n			ɲ	ŋ	
	l					
	ɬ					
	r					
w		j				

Vowels:

i		u	iː	yː	uː	ĩ	ỹ	ũ
e	ə	o	eː	øː	oː	ẽ	ø̃	õ
			ɛː			ɛ̃		
	a			aː			ã	

Final consonants:

Ø		
p		ʔ
m	n	
	r	
w		

The information capacity of each position in consonant clusters (conditioned by the preceding position) can be determined in the way shown above. Since there seem to be no important restrictions on how initial and final clusters combine with vowels in Classical Tibetan, their information capacity can be computed independently of the consonants.

In modern Tibetan there are no consonant clusters, but there are cooccurrence restrictions between vowels and final consonants, as shown in Table 5:2. The 2 bits of information due to the four tones can be distributed over the rhyme according to the principles given above. The resulting information distribution is shown in Table 6 and Figure 2.

The code shift is much more dramatic in Tibetan than in Chinese: the information contents of the vowel increases from 15% in Classical Tibetan to 50% in Modern Lhasa. Contrary to what happened in Chinese, the information contents of one phoneme increased considerably, from 3.7 to 5.7 bits/phoneme.

Table 5:1. Initial consonant clusters in Classical Tibetan.

	p	pʰ	b	m	t	tʰ	d	n	ts	tsʰ	dz	c	cʰ	ɟ	ɲ
Ø	+jr	+jr	+jr	+jr	+	+	+r	+jr	+	+w	+	+	+	+	+w
b					+		+		+			+			
d	+jr		+jr	+j											
g					+		+	+	+			+			+
m						+	+	+		+	+		+		+
ɦ		+jr	+jr			+	+r			+	+		+	+	
r			+	+j	+		+	+	+w		+			+	+
br					+		+	+	+		+			+	+
l	+		+		+		+					+		+	
bl					+		+								
s	+jr		+jr	+jr	+		+	+	+						+
bs					+		+	+	+						+

	k	kʰ	g	ŋ	s	z	ɕ	ʑ	r	l	w	j	h	ɦ	ʔ
Ø	+wjrl	+wjr	+jr	+	+	+l	+w	+w	+w	+	+	+	+w	+	+
b	+jrl		+jr		+	+l	+	+		+					
d	+jr		+jr	+							+				
g					+	+	+	+	+w	+		+			
m		+jr	+jr	+											
ɦ		+jr	+jr			+	+r								
r	+j		+j	+						+					
br	+j		+j	+						+					
l	+		+	+							+		+		
s	+jr		+jr	+					+	+					
bs	+jr		+jr	+					+	+					

Table 5:2. Combinations of vowels and final consonants in Modern Tibetan.

	Ø	p	ʔ	m	n	r	w
i	+	+	+	+		+	+
e	+	+	+	+		+	
ə		+					+
a	+	+	+	+	+	+	+
o	+	+	+	+	+	+	
u	+	+	+	+	+	+	

Long and nasal vowels occur only in open syllables.

The main historical processes that caused the code shift are loss of consonant clusters and of final consonants, accompanied by the development of tones, and of both length and nasality contrasts in the vowel system.

Table 6. Information distribution in Tibetan syllables.

Classical Tibetan:

Position	Length	Information bits	%
C_1	1	4.907	33
C_2	0.480	2.263	15
C_3	0.312	1.280	9
C_4	0.095	0.363	2
V	1	2.322	15
C_{f1}	0.900	3.322	22
C_{f2}	0.222	0.531	4
Total	4.009	14.988	

I/L = 3.739

Modern Tibetan:

Position	Length	Information segm.	tone		total	%
C_i	0.967				4.907	39
V	1	4.459	+ 1.730	=	6.189	50
C_f	0.234	1.126	+ 0.270	=	1.396	11
Total	2.201				12.492	

I/L = 5.676

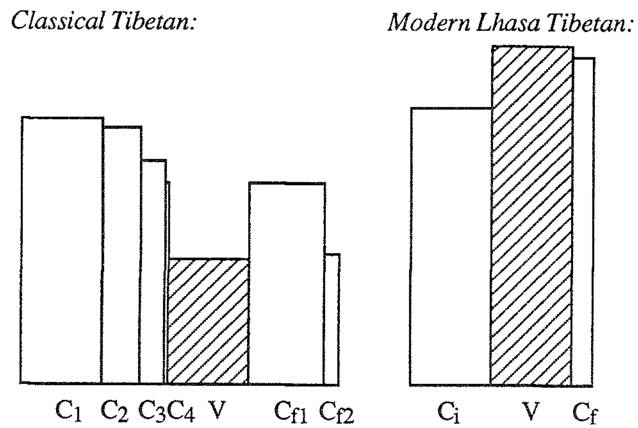

Figure 2. Information distribution in Tibetan syllables.

KAMMU

In many Austroasiatic languages of Southeast Asia, registers (i.e. voice quality contrasts) or tones have developed in conjunction with loss of the voicing contrast in initial consonants. Since registers and tones are manifested acoustically on the vowel (and perhaps on voiced final consonant as well), this is a change which transfers information from the initial consonant to the vowel. I will illustrate its effect in Kammu, an Austroasiatic language spoken mainly in Northern Laos. Proto-Kammu lacked tones, but some dialects, here labelled 'Northern Kammu' have developed a contrast between two tones, high and low, which replaced the original contrast between voiceless and voiced initial consonants. Some dialects ('Southern Kammu') have retained the original state without tones (see Lindell, Svantesson and Tayanin 1981, Svantesson 1983).

The segmental phonemes in Proto-Kammu and modern Northern Kammu are shown in Table 7. I will restrict myself to monosyllabic words. There are also a large number of sesquisyllabic words, consisting of a major syllable preceded by an unstressed minor syllable without a phonemic vowel, but minor syllables have not changed in the development from Proto-Kammu to modern Kammu, so they are not considered here.

The vowels and final consonants are assumed to have been the same in Proto-Kammu as in Northern Kammu (although minor changes may have taken place).

The initial consonant clusters in Kammu are shown in Table 8. In addition, there is one three-member cluster, *khw-*, which occurs in Lao loans and will be disregarded here.

There are also restrictions on the combination of vowels and final consonants, the most important ones being that long vowels do not combine with final *-ʔ* and *-h* (except *ɔɔ* and *εε*, which are phonetically shortened in this position), and that short vowels do not occur in open syllables.

Taking these restrictions into account, the information distribution over Kammu monosyllables can be computed as shown in Table 9 and Figure 3. The relative information contents of the vowel increased from 31% in Proto-Kammu to 35% in Northern Kammu, an effect which is entirely due to Northern Kammu tonogenesis.

The total information capacity of a monosyllable became slightly higher (about 2% increase). In a sense, tonogenesis frees the features involved from their segmental dependence. In Kammu the feature 'voiced/voiceless' in consonants was replaced with the tone feature 'low/high'. For instance, since Proto-Kammu did not have **z-*, there are no inherited words with initial *s-* and low tone. Nevertheless, a few such words have been introduced by borrowing, and it seems easier for speakers to accept new combinations of existing segments and

Table 7. Kammu segmental phonemes.

Initial consonants:

Proto-Kammu

p	t	c	k	ʔ
b	d	ɟ	g	
	s			h
m	n	ɲ	ŋ	
m̥	n̥	ɲ̥	ŋ̊	
ʔm	ʔn			
	l			
	l̥			
	r			
	r̥			
w		j		
w̥		j̥		
ʔw		ʔj		

Northern Kammu

p	t	c	k	ʔ
ɓ	ɗ			
	s			h
m	n	ɲ	ŋ	
	l			
	r			
w		j		
ʔw		ʔj		

Vowels:

i	ɨ	u	iː	ɨː	uː	ia	ɨa	ua
e	ə	o	eː	əː	oː			
	a		εː	ʌː	ɔː			
				aː				

Final consonants:

Ø

p	t	c	k	ʔ
	s			h
m	n	ɲ	ŋ	
	l			
	r			
w		j		

Table 8. Initial consonant clusters in Proto-Kammu. The clusters beginning with voiceless stops are found in Northern Kammu as well.

	l	r	w	h
p	+	+		+
t		+		+
c		+		+
k	+	+	+	+
b	+	+		
d		+		
ɟ		+		
g	+	+	+	

Table 9. Information distribution in Kammu monosyllables.

Proto-Kammu:

Position	Length	Information bits	%
C_1	1	4.954	36
C_2	0.147	0.661	5
V	1	4.322	31
C_f	0.956	3.892	28
Total	3.103	13.829	

I/L = 4.457

Modern Northern Kammu:

Position	Length	Information segm.	tone	total	%
C_1	1			4.248	30
C_2	0.152			0.659	5
V	1	4.322	+ 0.662	= 4.984	35
C_f	0.956	3.892	+ 0.338	= 4.230	30
Total	3.108			14.121	

I/L = 4.543

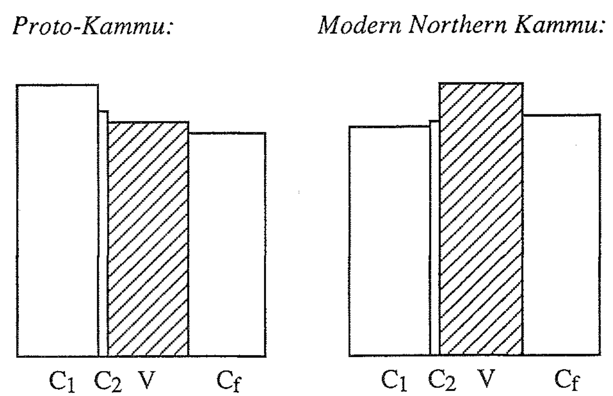

Figure 3. Information distribution in Kammu monosyllables.

existing tones than to introduce new segments. Similar cases are found in Chinese. In this way, the combinatorical possibilities of a language increases (although marginally) through tonogenesis.

MONGOLIAN

The Altaic languages differ radically from the three languages treated above, by having di- or polysyllabic morphemes and a rich suffixing morphology. Nevertheless, at least Mongolian, the Altaic language which has had the most extensive contacts with Chinese, has undergone a code shift of the East Asian type. In order to show this, I will compare Classical Mongolian, the written language originating in the 13th Century, and modern Khalkha, the standard language in the Mongolian People's Republic.

For illustration, I will use words with the structure (C)VCV, which is a common morpheme type in Mongolian.

The phonemes of the two languages are shown in Table 10. One important feature in Mongolian phonology is vowel harmony. The harmony classes are palatal (*e, y, ø*) vs. non-palatal (*a, u, o*) in Classical Mongolian, and pharyngeal (*a, ɷ, ɔ*, etc.) vs. non-pharyngeal (*e, u, ɵ*, etc.) in Khalkha. The vowel *i* is 'neutral' in both languages. There is also rounding harmony in Khalkha (see Svantesson 1985 for details). Because of vowel harmony, there are much fewer constrasts in the second vowel than in the first one, and the method of computing information must be adapted to cope with this discontinuous interdependency.

In Khalkha there are more consonant phonemes in words with pharyngeal vowels than in non-pharyngeal words. Consonants within parentheses in Table 10 occur only in pharyngeal words. Since there are, (in addition to Ø), 23 different initial consonants in pharyngeal and 17 in non-pharyngeal words, the information contents of the initial consonant can be given as log ((24 + 18)/2) = 4.392 bits. The information of a Classical Mongolian initial consonant is computed straightforwardly to log 15 = 3.907 bits. The information of the medial consonant is computed in a similar way.

All vowels (7 in Classical Mongolian, 36 in Khalkha) contrast in the first vowel position, but given the first vowel, the vowel in the second position is restricted by harmony, as shown in Table 11. The information contents of the second vowel must be computed by conditioning on the first vowel. For example, in Classical Mongolian, it is log ((1·5 + 1·4 + 5·3)/7) = 1.778 bits. The information distribution is shown in Table 12 and Figure 4.

As seen there, the information carried by the vowels increased from 36% in Classical to 49% in Khalkha Mongolian. Furthermore, the total information capacity of a CVCV structure increased considerably, so the phonological

Table 10. Mongolian segmental phonemes.

Initial consonants:

Classical Mongolian				Modern Khalkha						
Ø				Ø						
p	t		k/q	p	(p´)	t				
b	d		g/ɢ	b	(b´)	d		(g´)	g/ɢ	
	s	ɕ				s	ɕ	(x´)	x	
		tɕ				ts	tɕ			
		dʑ				dz	dʑ			
m	n			m	(m´)	n	(n´)			
	l					ɮ				
		j		w		j				

Vowels:

Classical Mongolian			Modern Khalkha							
i	y	u	i		u	iː		uː		ui
e	ø	o	e		ɷ	eː		ɷː		ɷi
		a			ɵ			ɵː		
				a	ɔ		aː	ɔː	ai	ɔi

For each oral vowel there is a corresponding nasal vowel phoneme.

Medial consonants:

Classical Mongolian				Modern Khalkha						
p	t		k/q	p	(p´)	t	(t´)			
b	d		g/ɢ			d	(d´)	(g´)	g	(ɢ)
	s	ɕ				s	ɕ	(x´)	x	
		tɕ				ts	tɕ			
		dʑ				dz	dʑ			
m	n			m	(m´)	n	(n´)			
	l					ɮ	(ɮ´)			
	r					r	(r´)			
		j		w	(w´)	j				

changes involved are not compensatory as in the languages treated above. Instead, both vowels and consonants increased their information capacity, but the increase was greatest for the vowels (as pointed out by Maddieson 1986, the sizes of vowel and consonant inventories are positively correlated in the world's languages, so compensatory changes cannot generally be expected).

The phonological processes involved in the Mongolian code shift are loss of non-initial short vowels, the development of nasal vowels from combinations of vowels and a nasal consonant, and the development of long vowels from certain CVC forms. Changes that increased the consonantal contrasts are palatalisation and the development of dental affricates.

Table 11. Possible combinations of the first and second vowel in Mongolian.

Classical Mongolian:

	i	e	y	a	o	u
i	+	+	+	+		+
e	+	+	+			
ø	+	+	+			
y	+	+	+			
a	+			+		+
o	+			+	+	+
u	+			+		+

Khalkha Mongolian:

	iː	eː	ɵː	uː ui	aː ai	ɔː ɔi	ɷː ɷi
i, iː	+	+		+			
e, eː	+	+		+			
ɵ, ɵː	+		+	+			
u, uː, ui	+	+		+			
a, aː, ai	+				+		+
ɔ, ɔː, ɔi	+					+	+
ɷ, ɷː, ɷi	+				+		+

Nasal vowels combine in the same way as the corresponding oral vowels. There is a nasal shwa /ɜ̃/, which can occur as the second vowel in combination with any first vowel.

These changes have also shortened the Mongolian words considerably, so that Khalkha CVCV-words often correspond to words with three or more syllables in Classical Mongolian.

In some Mongolian dialects spoken in Inner Mongolia in Northern China, palatalised consonants have umlauted the vowels and then merged with the corresponding plain consonant, a process which transferred still more information to the vowels.

CONCLUSION

The method of computing the information distribution over syllables (or other relevant units) developed here has revealed a common feature of the diverse historical phonological developments that languages in the East Asian area have undergone. This common feature is a transfer of relative information capacity from consonants to vowels, which can be regarded as a code shift in the sense of Sigurd 1961.

Table 12. Information distribution in Mongolian CVCV-words.

Classical Mongolian:

Position	Length	Information bits	%
C_1	0.933	3.907	31
V_1	1	2.807	22
C_2	1	4.000	32
V_2	1	1.778	14
Total	3.933	12.492	

I/L = 3.176

Khalkha Mongolian:

Position	Length	Information bits	%
C_1	0.951	4.392	25
V_1	1	5.170	30
C_2	1	4.585	26
V_2	1	3.322	19
Total	3.951	17.469	

I/L = 4.421

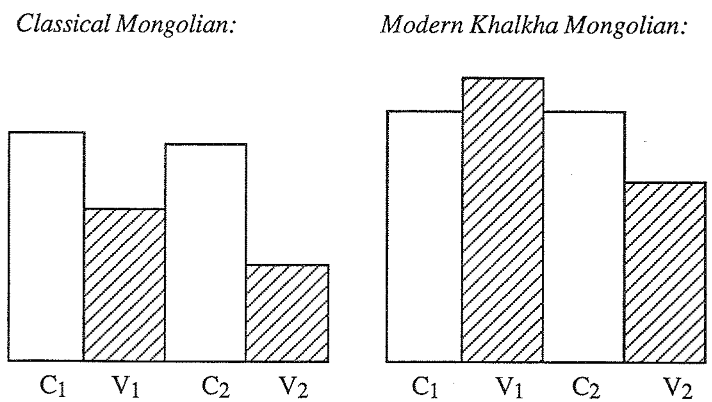

Figure 4. Information distribution in Mongolian CVCV-words.

Languages from several language families have undergone this East Asian code shift – it was exemplified here by Chinese, Tibetan, Kammu and Mongolian, which represent all the major language families in East Asia, except Kadai (but comparison of Proto-Tai (Li 1977) with modern languages such as Thai or Lao makes it clear that the code shift has taken place in this family as well).

The East Asian code shift is not the result of a single phonological process, but each language seems to have chosen its own ad hoc methods to reach the common goal of concentrating information on the vowels. It seems difficult to explain this joint effort by appeal to language internal factors such as economy principles (Martinet 1952) or preference laws (Vennemann 1988). Instead, the East Asian code shift is another demonstration of the importance of areal trends in historical linguistics.

REFERENCES

Ash, Robert. 1965. *Information theory.* New York: Interscience.

Billingsley, Patrick. 1965. *Ergodic theory and information.* New York: John Wiley.

Chen, Matthew. 1976. 'From Middle Chinese to modern Peking'. *Journal of Chinese Linguistics* 4, 113-277.

Fù Jiāzhāng. 1983. *Pô-gyǎ šǎnčāa ki Lhǎsā khakà čhīngtsò / Zàng-Hàn duìzhào Lāsà kǒuyǔ cídiǎn.* Běijīng: Mínzú chūbǎnshè.

Jaglom, Akiva M. and Isaak M. Jaglom. 1973. *Verojatnost' i informacija.* 3rd ed. Moskva: Nauka.

Li Fang Kuei. 1977. *A handbook of comparative Tai.* Honolulu: University Press of Hawaii.

Lindell, Kristina, Jan-Olof Svantesson and Damrong Tayanin. 1981. 'Phonology of Kammu dialects'. *Cahiers de linguistique Asie Orientale* 9, 45-71.

Maddieson, Ian. 1986. 'The size and structure of phonological inventories: analysis of UPSID'. *Experimental Phonology,* ed. John Ohala and Jeri Jaeger, 105-23. Orlando: Academic Press.

Martinet, André. 1952. 'Function, structure, and sound change'. *Word* 8, 1-32.

Meyer-Eppler, W. 1959. *Grundlagen und Anwendungen der Informationstheorie.* Berlin: Springer-Verlag.

Shannon, Claude. 1948. 'A mathematical theory of communication'. *Bell System Technical Journal* 27, 379-423, 623-56. Reprinted 1949 in *The mathematical theory of communication,* by Claude Shannon and Warren Weaver, 31-125. Urbana: University of Illinois Press.

Sherard, Michael. 1980. *A synchronic phonology of modern colloquial Shanghai (=Computational Analyses of Asian & African Languages* 15). Tōkyō: National Inter-University Research Institute of Asian and African Languages and Cultures.
Sigurd, Bengt. 1961. 'The code shift in Old Norse'. *Studia Linguistica* 15, 10-21.
Sigurd, Bengt. 1965. *Phonotactic structures in Swedish.* Lund: Uniskol.
Svantesson, Jan-Olof. 1983. *Kammu phonology and morphology.* Lund: Gleerup.
Svantesson, Jan-Olof. 1985. 'Vowel harmony shift in Mongolian'. *Lingua* 67, 283-327.
Vennemann, Theo. 1988. *Preference laws for syllable structure and the explanation of sound change.* Berlin: Mouton de Gruyter.

Lund University, Dept. of Linguistics
Working Papers 33 (1988), 215-227

What else Do you Need to Know about the Referents?

Karina Vamling

Abstract
In a discourse fragment, the referents are referred to by various noun phrases, pronouns and personal markers in the verbform. It is interesting to consider the substitutional character of the personal pronouns. What information do they express about the referents, i. e. what does one need to know about a referent in order to pick the right pronoun?

In a case like *I know Johan. He works at our department*, the pronoun *he* is chosen for the second reference to the person called *Johan,* as it corresponds to one male person, who is neither the speaker nor the hearer. The features involved are thus, person, number and semantic gender.

The selection of pronouns is somewhat different in, for instance, the Caucasian languages, where additional features of the discourse situation become relevant, such as the distance and relative location of the referents above or below the speaker/hearer.

The implications such systems may have for a formal, cross-linguistic model are briefly discussed in the paper.

BACKGROUND

Before going into detail in the system of personal and demonstrative pronouns in the Caucasian languages, some general points about systems of personal pronouns will be noted.

An early proposal about pronoun systems was made by Greenberg 1963, stating that "All languages have pronominal categories involving at least three persons and two numbers" (p. 96). A following study by Ingram 1978 proposes refinement of Greenberg's universal: "every language designates at least four persons – I, thou, he and we" (p. 227). Ingram also states that the minimal distinction that is made in number systems is between 'one' and 'more than one'. Ingram bases his generalizations on a sample of approximately 60 languages.

A larger study the typology of personal pronoun systems has been conducted by Sokolovskaja 1980, who takes as her basis personal pronouns in 400 languages. She formulates the generalization: "Each of the seven meta-persons[1] is expressed in the system of personal pronouns in any language" (p. 90). She

[1]The metapersons are defined as: я=I, ты=thou, он=he: 1. (I)^X I, 2. (you)^X thou, you, 3. (he)^X he, they, 4. (I & you)^X we incl., 5. (I & he)^X we excl., 6. (you & he)^X you, 7. (I & you & he)^X we [K.V.].