FREQUENCY OF OCCURRENCE OF PHONEMES AND SYLLABLES IN THAI: ANALYSIS OF SPOKEN AND WRITTEN CORPORA

A. Munthuli[1], C. Tantibundhit[1, 2], C. Onsuwan[2, 3], K. Kosawat[4], C. Wutiwiwatchai[4]

[1]Department of Electrical and Computer Engineering, Thammasat University, Thailand
[2]Center of Excellence in Intelligent Informatics, Speech and Language Technology and Service Innovation (CILS), Thammasat University, Thailand
[3]Department of Linguistics, Thammasat University, Thailand
[4]National Electronics and Computer Technology Center (NECTEC), Thailand
tchartur@engr.tu.ac.th, consuwan@tu.ac.th, krit.kosawat@nectec.or.th, chai.wutiwiwatchai@nectec.or.th

ABSTRACT

This work provides detailed frequency and distribution of Thai phonemes, biphones, and syllable types drawn from three large-scale Thai corpora (InterBEST, LOTUS-BN, and LOTUS-Cell 2.0). Comparisons are carried out to examine an extent to which linguistic variation, associated with different corpus types (written vs. spoken), affects frequency statistics and distribution patterns. Results and statistical analysis show that there is a high correlation in terms of occurrence frequency and distribution in the case of tones and syllable types. However, large degrees of discrepancy exist among the data sets of initial consonants, vowels, and final consonants. Comparisons of this type are needed for other languages to reliably show the degrees to which different types of language corpus and linguistic variation contribute to variability in phoneme frequency and distribution.

Keywords: phoneme frequency, biphone, syllable structure, Thai, corpus analysis

1. BACKGROUND

Knowledge and understanding of phoneme occurrence and distribution are essential for research in all areas of speech technology. Moreover, they are relevant for linguistics research, language teaching, and clinical domain, such as language pathology and speech audiometry [1]. For instance, measuring hearing thresholds in speech audiometry partly relies on the use of phonetically or phonemically balanced (PB) word lists that truly reflect language's distribution of phonemes [2].

Current studies on phoneme frequency have been carried out with the aid of large language database, i.e., language corpora [3]. However, not much work has directly addressed the question of whether well-known linguistics variation, (e.g., different lexical variety) that comes from different types of corpus, may contribute to variability in the frequency and distribution patterns. Sandoval *et al.* observed different frequency results between written and spoken corpora of Castilian Spanish, especially for certain types of vowels [4]. However, no statistical analysis was performed to confirm if these are statistically significant.

Munthuli *et al.* was among the first studies on the Thai language that successfully obtained frequency and distribution of phonemes from a large-scale written corpus (InterBEST) [5]. Bearing in mind the issue of the linguistic variation, the present study takes further steps and compares frequency statistics and distribution patterns across three large-scale Thai corpora (one written and two spoken) by using the same reliable criterion and tools.

2. THAI CORPUS

Table 1 summarizes main characteristics of the three Thai corpora used in the present study. The novelty of this work is the use of large-scale corpora which are highly reliable. Out of the three corpora, one (InterBEST) is written (text-based) and the other two (LOTUS-BN and LOTUS-CELL 2.0) are spoken (speech). The spoken databases include read speech from broadcast news and spontaneous speech from telephone conversation. It should be noted that Named Entities (NEs) (e.g., acronyms, abbreviations, and foreign names that are obviously 'thaified'), which constitute a large portion in the databases, are present in our analysis.

Table 1: Characteristics of InterBEST (BEST), LOTUS-BN (LT-BN), and LOTUS-Cell 2.0 (Casual (LT-CS) and Formal styles (LT-FS)).

	BEST	LT-BN	LT-CS	LT-FS
Types	Written	Spoken	Spoken	Spoken
Words	6,969,608	829,494	476,733	477,925
Unique words	94,729	35,172	9,425	9,504
Syllables	9,996,343	1,230,207	568,721	573,221

2.1. InterBEST (BEST)

InterBEST is one of the largest Thai written corpora that are publicly available. It is composed of 12 text genres (e.g., encyclopaedia, novels, and news articles) amounting to approximately 9 million words [6]. To be in line with [5], data from three text genres (Law, National Software Contest report, Royal news), whose phoneme distribution did not fall within the 95% confidence interval, were excluded. At least 80% of the data, approximately 7 million words, remained [5].

2.2. LOTUS-BN (LT-BN)

LOTUS-BN is a Thai spoken corpus, drawn from Thai television broadcast news in 18 topics (e.g., crime, weather report, and politics) [7]. It includes approximately 100 hours of audio recordings from 43 female and 38 male speakers.

2.3. LOTUS-Cell 2.0 (LT-FS and LT-CS)

LOTUS-Cell 2.0 is one of the largest Thai telephone conversation corpora, with recordings of approximately 50 hours of monologues (responses to questions) and (turn-taking) conversations from 213 speakers [8]. It was designed for the use of automatic speech recognition system training [8]. The corpus contains many interesting and useful annotations which include 'filled' and 'unfilled' pauses, and variations in pronunciation (formal and casual styles). Transcription of real speech is not a straightforward task as it possesses many unique characteristics that include variations in lexical items and pronunciation. Interestingly, LOTUS-Cell 2.0's annotations were designed to draw a distinction between formal (LT-FS) and casual speech styles (LT-CS). LT-CS reflects the real use of colloquial speech as much as possible. For example, it includes sentence particles and filled pauses (i.e., [náʔ], [ʔūːm]) and allophonic variants (i.e., the alternation of [r] and [l] and /r/ and /l/ deletion in consonant clusters), which are annotated as 'incorrect pronunciations' [8]. On the other hand, the formal ('correct and standard') forms are given and represented in LT-FS.

3. ANALYSIS

It should be noted that all three corpora are kept in the form of Thai graphemes and related annotations (tagging). Desired frequency output is derived by means of automatic software (grapheme to phoneme conversion (G2P)). Therefore, similar tool and technique are applied to each corpus. The major steps include:

1) Pre-processing: Extract all annotations, separate LT-FS from LT-CS, perform word boundary parsing, and identify unique words.
2) Grapheme to phoneme conversion: Apply Vaja 6.0 [9] to transcribe unique words into phoneme description, e.g.,"เหมือน" is transcribed to "m-vva-n^-4|*". Note that the G2P [9] has an approximate accuracy of 88.75% such that some errors may present in the output. In the present study, all transcribable words by Vaja 6.0 are used. The non-transcribable words from Vaja 6.0 output were manually transcribed by a well-trained student, verified by a linguist, and added back to the corpus.
3) Occurrence frequency calculation: Calculate frequency of occurrence of unique words from step 1) and count number of occurrences of each phoneme and group phonemes into initials, finals, vowels, and tones.

3.1. Statistical analysis

To investigate dependency between two categorical variables, i.e., corpus vs. phoneme-type and corpus vs. syllable-type, corpus is considered an independent variable represented in row, while phoneme-type and syllable-type a dependent variable represented in column. Contingency table chi-square test is performed to reflect the strength of relationship between variables at 0.05 level of significance. We hypothesize that the proportion of each individual phoneme, e.g., /p/ in one phoneme-type (initial) across four corpora is equally likely. It is important to note that every cell frequency is a relative frequency value. Fisher's exact test [10] is performed in cases where there are more than 20% of cell frequencies whose values are under five (the cases of initials, finals, and vowels).

3.2 Thai Phonology

To interpret the results, a brief description of the phonology of Thai and symbols used in the analysis should be given. Thai syllables may be represented as $C_i(C)V^TC_f$ or $C_i(C)V{:}^T$, where C_i stands for an initial consonant, C_iC a consonantal cluster, C_f a final consonant, V a short vowel, V: a long vowel, and T a tone. Phonologically, a Thai syllable never starts without a consonant and always bears a tone. It is important to note that we mostly adopt phonemic description of Thai proposed by Tingsabadh and Abramson with some modifications [11]. Unlike them, we draw a distinction between

short and long diphthongs. /w/ and /j/ after a vowel, which are treated here as final consonants, phonetically are semivowels [12].

Importantly, some types of allophonic variation are represented in our analysis, especially alternation of [r] and [l] and /r/ and /l/ deletion in consonant clusters. However, others, such as tone neutralization (low and high tones into mid tone) and vowel shortening are not represented. It is well known that /r/ and /l/ distinction has become an unstable one for many Thais [11]. In real speech, /r/ has gradually turned into [l] (/l/ exists as a separate phoneme), and also /r/ and /l/ are often deleted in initial clusters. As these phenomena are one of the important characteristics of real (colloquial) speech, we decided to keep them in the analysis (in LOTUS-Cell 2.0).

Another important issue is representation of a final glottal stop /ʔ/. Phonemically and when spoken in isolation, a short-vowel monosyllabic word without any other final consonant is ended with a glottal stop. However, short-vowel syllables in polysyllabic words and short-vowel monosyllabic words in continuous speech (between pauses) are often pronounced with no final glottal stop. To resolve this issue and to make the analysis as consistent as possible, we decided to use 'x' to signify an ending of any short-vowel syllables with no final consonant. On the other hand, an ending of a syllable with long vowel with no final consonant is represented as 'ø'.

4. RESULTS

4.1. Initial Consonant

A discrepancy is observed in many cases, such as /tʰ/ /r/ /l/ /ʔ/ /kʰr/. In fact, Fisher's exact test gives p-value $= 3.37 \times 10^{-7}$ and null hypothesis is rejected. Despite the differences, we found that common top-5 phonemes are /s/ and /k/. The discrepancy could perhaps be attributed to the alternation of [r] and [l] and /r/ and /l/ deletion in consonant clusters, which are important characteristics of real (colloquial) speech (as previously mentioned). Moreover, /ʔ/ occurs much more frequently in LT-CS and LT-FS. This could be due to the fact that many 'filled pauses' in continuous speech in Thai start with this phoneme (e.g., [ʔə̄ː], [ʔtūːm]).

4.2. Vowel

Like initial consonants, many differences are found in the percentage of frequency and in the rank order of Thai vowels across the corpora. Fisher's exact test gives p-value $= 3.94 \times 10^{-6}$ and null hypothesis is rejected. Despite the differences, common top-5 vowels are /a/ /aː/ /ɔː/ and /iː/. These four vowels amount to approximately 63% of all vowel occurrences. Interestingly, out of all vowel occurrences, 52.72% are long vowel (47.28% are short).

4.3. Final consonant

Like initial consonants and vowels, differences can be seen in the percentage of frequency and in the rank order of final consonants across the corpora. Fisher's exact test gives p-value $= 9.64 \times 10^{-7}$ and null hypothesis is rejected. Despite the differences, common top-5 final consonants are /ŋ/ 'x' /n/ /j/ and /t/. Particularly, /ŋ/ 'x' /n/ and /j/ account for more than 50% of all final consonant occurrences.

4.4. Lexical tone

There is a large agreement in the percentage of frequency and in the rank order of lexical tones across the corpora. [$\chi^2(12) = 1.526, p = 0.9999$] and null hypothesis fails to reject. Mid tone occurs with the highest frequency, followed by low tone (falling tone in LT-CS and LT-FS), and falling tone (low tone in LT-CS and LT-FS). Mid and low tones amount to more than 50 % of all tones.

4.5. Syllable type

Like lexical tones, percentage of frequency and rank order of syllable types across the corpora follow highly similar patterns [$\chi^2(9) = 1.934, p = 0.9925$] and null hypothesis fails to reject. CVC occurs with the highest frequency, followed by CVVC (CVVø in LT-CS and LT-FS), CVVø (CVVC in LT-CS and LT-FS), and CVx with the lowest frequency. Closed syllables (CVC and CVVC) constitute nearly 60% of all syllables.

Table2: Percentage of frequency of occurrence of initial consonants in descending order (ordered by BEST results).

	s	tʰ	n	m	k	kʰ	r	l	pʰ	t	d	tɕ	j	p	w	tɕʰ	ʔ	h	b
BEST	8.9	8.2	7.5	6.6	6.5	6.2	6.1	4.8	4.7	4.1	4.0	3.9	3.7	3.5	3.2	3.1	3.0	3.0	2.3
LT-BN	8.7	8.0	9.4	5.1	6.9	7.1	5.9	4.9	4.1	4.1	3.7	3.7	2.9	2.9	3.6	3.3	3.0	2.7	2.3
LT-CS	7.1	5.8	6.6	7.3	7.6	8.9	1.2	10.2	4.1	3.4	4.2	3.6	3.4	4.4	3.5	2.8	7.5	2.6	2.5
LT-FS	7.1	5.8	7.1	7.2	7.0	6.7	6.3	6.4	3.5	3.2	4.0	3.6	3.9	3.6	3.5	2.7	6.6	2.6	2.5

Table 2 (continued): Percentage of frequency of occurrence of initial consonants in descending order.

	pr	kʰw	pʰr	kʰr	kl	ng	kr	f	tr	pl	kw	kʰl	pʰl	tʰr
BEST	1.2	0.8	0.8	0.7	0.6	0.6	0.5	0.5	0.4	0.3	0.2	0.2	0.1	~0.0
LT-BN	1.4	0.5	0.7	2.0	0.5	0.6	0.6	0.6	0.5	0.2	0.1	0.2	0.1	~0.0
LT-CS	0.1	0.3	0.1	0.1	0.1	1.3	0.1	0.6	0.0	0.0	0.2	0.0	0.1	~0.0
LT-FS	0.5	0.3	0.6	2.3	0.3	0.5	0.4	0.6	0.2	0.4	0.2	0.1	0.2	~0.0

Table 3: Percentage of frequency of occurrence of vowels in descending order (ordered by BEST results).

Table 4: Percentage of frequency of occurrence of final consonants in descending order (ordered by BEST results).(Descriptions of 'x' and 'ø', see text).

	a	aː	ɔː	iː	i	o	u	ɛː	e	uaː	uː	ɯ
BEST	28.7	21.1	7.0	6.6	5.0	5.0	3.1	2.8	2.6	2.4	2.3	1.9
LT-BN	30.7	20.1	7.2	7.4	4.7	4.6	2.9	2.6	2.0	2.5	2.1	1.6
LT-CS	30.7	17.0	9.8	6.5	3.6	3.4	2.4	4.4	2.6	1.7	2.2	1.6
LT-FS	30.5	17.5	9.3	7.3	3.6	3.4	2.4	4.4	2.6	1.7	2.2	1.5
	ɯaː	eː	ɤː	oː	ɯː	iaː	ɛ	ɔ	ɤ	ia	ua	ɯa
BEST	1.8	1.7	1.4	1.4	1.4	1.4	1.4	0.8	0.1	~0.0	~0.0	~0.0
LT-BN	1.6	2.3	1.6	1.7	1.1	1.4	1.2	0.5	0.2	~0.0	~0.0	~0.0
LT-CS	1.5	1.9	2.6	1.5	2.4	2.2	0.7	1.0	0.3	~0.0	~0.0	~0.0
LT-FS	1.5	1.9	2.5	1.5	3.0	1.4	0.7	0.9	0.3	~0.0	~0.0	~0.0

	'ø'	n	'x'	ŋ	j
BEST	21.4	15.5	13.7	11.4	10.9
LT-BN	23.0	14.6	16.1	10.3	8.8
LT-CS	27.6	13.3	13.5	8.8	12.2
LT-FS	27.2	13.2	13.2	9.3	12.3
	t	m	k	p	w
BEST	7.8	6.6	5.4	4.0	3.2
LT-BN	8.5	5.9	4.8	5.2	2.7
LT-CS	5.1	4.7	4.3	5.1	5.2
LT-FS	5.2	4.7	4.3	5.1	5.4

Table 5: Percentage of frequency of occurrence of lexical tones in descending order (ordered by BEST results).

	mid	low	falling	high	rising
BEST	33.7	22.3	19.2	15.3	9.4
LT-BN	33.6	22.0	18.6	18.4	7.5
LT-CS	32.3	20.8	22.5	15.6	8.8
LT-FS	31.7	20.4	22.2	15.9	9.8

4.6. Biphone

Table 7 shows common top-50 biphones (pairs of phonemes) out of more than 1,000 existing pairs. Twenty-one out of 37 pairs are VCs and the rest are CVs. In the analysis, a consonant cluster and a diphthong are treated as a single unit.

Table 7: Common (top-50) biphones in BEST, LT-BN, LT-CS and LT-FS. Tones are not analyzed.

CV biphone pairs	VC biphone pairs
[ma] [na] [kʰa] [tɕa] [sa] [tʰiː] [pa] [waː] [maː] [tʰa] [ka] [daː] [tɕʰa] [kaː] [saː] [kɔː]	[ax] [aj] [iːø] [aːø] [ɔːø] [an] [ap] [aːŋ] [aːj] [aw] [aːn] [en] [ɔːŋ] [aŋ] [uːø] [on] [am] [ɯŋ] [aːm] [it] [aːk]

Table 6: Percentage of frequency of occurrence of syllable types in descending order (ordered by BEST results). (Descriptions of 'x' and 'ø', see text).

	CVC	CVVC	CVVø	CVx
BEST	34.9	29.9	21.4	13.7
LT-BN	32.4	28.6	23.0	16.1
LT-CS	32.7	26.2	27.6	13.5
LT-FS	32.6	27.0	27.2	13.2

4. DISCUSSIONS

We believe that we successfully carried out a detailed analysis of frequency and distribution of Thai phonemes and syllables. Importantly, comparisons among the data from three large-scale corpora were systematically performed. The findings show that there is a high correlation in terms of occurrence frequency and distribution in the case of tones and syllable types. However, large degrees of discrepancy exist among the data sets of initial consonants, vowels, and final consonants.

It should be useful to perform separate pair-wise statistical analysis among the four databases (i.e., BEST and LT-BN, LT-FS and LT-CS). Comparisons similar to the ones carried out here are needed for other languages to reliably show the degrees to which different types of language corpus and linguistic variation contribute to variability in language's phoneme frequency and distribution.

6. REFERENCES

[1] Mines, M.A., Hanson, B.F., and Shoup, J.E., Frequency of occurrence of phonemes in conversational English, *Language & Speech,* vol. 21, pp. 221-241, 1978.

[2] Tillman, T.W., and Carhart,. R., An expanded test for speech discrimination utilizing CNC monosyllabic words, USAF school of aerospace medicine, aerospace medical division (AFSC), Brooks air force base, TX1966.

[3] Hammer, A., *et al.*, Balancing word lists in speech audiometry through large spoken language corpora, in *Proc. 14th Annu. Conf. of the Int. Speech Commun. Assoc.*, 2013, pp. 3613-3616.

[4] Sandoval, A.M., *et al.*, Developing a Phonemic and Syllabic Frequency Inventory for Spontaneous Spoken Castilian Spanish and their Comparison to Text-Based Inventories, in *Proc. 6th Int. Conf. on Language Resources and Evaluation*, 2008, pp. 1097-1100.

[5] Munthuli A., *et al.*, A corpus-based study of phoneme distribution in Thai, in *Proc. 10th Int. Symp. on Natural Language Process.*, Phuket, TH, 2013, pp. 114-121.

[6] Kosawat, K., *et al.*, BEST 2009: Thai word segmentation software contest, in *Proc. 8th Int. Symp. on Natural Language Process.*,, Bangkok, TH, 2009, pp. 83-88.

[7] Chotimongkol, A., *et al.*, LOTUS-BN: A Thai Broadcast News Corpus and Its Research Applications, in *Proc. 12th Oriental Chapter of the Int. Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (Oriental COCOSDA)*, Xinjiang, CHN, 2009.

[8] Chotimongkol, A., *et al.*, The Development of a Large Thai Telephone Speech Corpus: LOTUS-Cell 2.0, in *Proc. 13th Oriental Chapter of the Int. Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (Oriental COCOSDA)*, Kathmandu, NP, 2010.

[9] Thangthai, A., *et al.*, Automatic syllable-pattern induction in statistical Thai text-to-phone transcription, in *Proc. 9th Int. Conf. on Spoken Language Process.*, Pittsburgh, US, 2006.

[10] Fisher, R.A., On the interpretation of $\chi 2$ from contingency tables, and the calculation of P, *J. of the Royal Statistic Soc.*, vol. 85, pp. 87-94, 1922.

[11] Tingsabadh, K., and Ambrason, A.S., Thai, *J. Int. Phonetic Assoc.* , vol. 20, pp. 24-28, 1993.

[12] Ambrason, A.S., The vowels and tones of standard Thai: Acoustical measurements and experiments. Indiana U. Research Center in Anthropology, Indiana University Research Center in Anthropology, Folklore and Linguistics., Bloomington, IN, 1962.