REPRESENTING TONE IN LEVENSHTEIN DISTANCE

CATHRYN YANG AND ANDY CASTRO

Abstract *Levenshtein distance, also known as string edit distance, has been shown to correlate strongly with both perceived distance and intelligibility in various Indo-European languages (Gooskens and Heeringa, 2004; Gooskens, 2006). We apply Levenshtein distance to dialect data from Bai (Allen, 2004), a Sino-Tibetan language, and Hongshuihe (HSH) Zhuang (Castro and Hansen, accepted), a Tai language. In applying Levenshtein distance to languages with contour tone systems, we ask the following questions: 1) How much variation in intelligibility can tone alone explain? and 2) Which representation of tone results in the Levenshtein distance that shows the strongest correlation with intelligibility test results? This research evaluates six representations of tone: onset, contour and offset; onset and contour only; contour and offset only; target approximation (Xu & Wang, 2001), autosegments of H and L, and Chao's (1930) pitch numbers. For both languages, the more fully explicit onset-contour-offset and onset-contour representations showed significantly stronger inverse correlations with intelligibility. This suggests that, for cross-dialectal listeners, the optimal representation of tone in Levenshtein distance should be at a phonetically explicit level and include information on both onset and contour.*

1. INTRODUCTION

The Levenshtein distance algorithm measures the phonetic distance between closely related language varieties by counting the cost of transforming the phonetic segment string of one cognate into another by means of insertions, deletions and substitutions. After Kessler (1995) first applied the algorithm to dialect data in Irish Gaelic, Heeringa (2004) showed that cluster analysis based on Levenshtein distances agreed remarkably with expert consensus on

International Journal of Humanities and Arts Computing 2 (1–2) 2008, 205–219
DOI: 10.3366/E1753854809000391

Dutch dialect groupings. In addition, Gooskens and Heeringa (2004) found a significant correlation between Levenshtein distance and perceived distance among Norwegian listeners ($r = .67$, $r < .001$), and Gooskens (2006) found an even stronger correlation with intelligibility among Scandinavian languages ($r = -.82$, $p < .001$).

Previously, the application of Levenshtein distance has been limited to Indo-European languages. However, Yang (accepted) showed that hierarchical clustering based on Levenshtein distance paralleled that of historical-comparative analysis in Nisu, a Tibeto-Burman language. Also, Levenshtein distance showed a strong, significant correlation with intelligibility test results ($r = -.62$, $p < .001$) in Nisu. This correlation suggests that Levenshtein distance is a good approximation of intelligibility for East Asian tonal languages. We apply Levenshtein distance to word lists from *Bai Dialect Survey* (Allen, 2004) and *Hongshuihe Zhuang Dialect Intelligibility Survey* (Castro and Hansen, accepted) and correlate the results with intelligibility test scores obtained during the respective surveys.

Contour tone is a distinguishing characteristic of many East Asian language families such as Sino-Tibetan and Tai-Kadai. Norwegian also has phonemic tone, but Gooskens and Heeringa (2006) found that prosody in Norwegian as measured by Levenshtein distance showed only a weak correlation with perceived distance ($r = .24$, $p < .01$) and could only explain 6 per cent of the variance. However, in their perceptual experiment with Chinese dialects, Tang and van Heuven (2007) compared natural speech recordings with recordings that had the pitch variations synthetically removed. They found that listeners made better subjective judgments about dialect distance with the fully tonal recordings than with the recordings that had the tonal information removed. In this paper, our first question investigates the relationship between tone and intelligibility: how much of the variation in intelligibility test results can be explained by tone alone? We measure tone and segment distance separately, correlate the distances with intelligibility test scores, then use multiple linear regression analysis to see which variable has the greatest relative contribution to intelligibility.

Additionally, we investigate the optimal way of representing tones in the Levenshtein distance algorithm in relation to intelligibility. We evaluate six representations of tone, although there are many more possible ways. We correlate Levenshtein distance using each tone representation with intelligibility to answer the question: Which way of representing tone in the Levenshtein distance algorithm shows the strongest correlation with intelligibility? Phonetically explicit representations, which include information on tonal onset, contour, and offset, are compared with more phonemic representations of tone as autosegments and targets. We also include Chao's pitch numbers, since this is the most widely used method of transcribing tone in East Asian tone languages.

2. MATERIAL

We used material from two dialect surveys on minority languages in China: Allen (2004) on the Bai language of Yunnan Province, and Castro and Hansen (accepted) on Hongshuihe (HSH) Zhuang in Guangxi Zhuang Autonomous Region. Both surveys included 500-item wordlists and intelligibility testing using Recorded Text Testing (RTT) methodology (see Section 3.3).

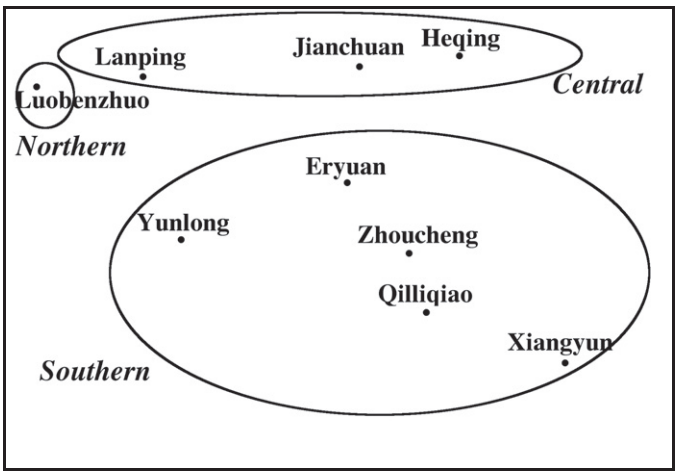
2.1 Bai

Bai is an ethnic minority language in southwest China with a population of around 1.8 million, with the vast majority located in Dali Bai Autonomous Prefecture (Allen, 2004; Yunnan Provincial Statistics Department, 2004). Although Bai is definitely Sino-Tibetan, linguists disagree as to whether it belongs in the Sinitic or the Tibeto-Burman branch of the family (Wang, 2005; Matisoff, 2003).

Previous research on Bai dialects group varieties into Central, Southern, and Northern (Xu and Zhao, 1984; Allen, 2004). Northern Bai, known to its speakers as Leimei, is considerably different from Central and Southern Bai, though still closely related (Bradley, 2007). Wang (2006) identifies a key tonal innovation that groups Southern varieties together: the development of a mid-rising tone in Proto-Bai *Tone 1a in syllables with *voiceless unaspirated and some *voiced sonorant initials. In other environments, *Tone 1a shows a high tone in Southern Bai dialects. Central and Northern Bai show a high tone in *Tone 1a in all environments. Also, Central and Southern Bai share the devoicing of *voiced stops, while Northern Bai retains a voiced stop series.

Allen (2004) collected 500-item word lists in nine locations, including varieties from each dialect group. Northern Bai was represented by Lushui Luobenzhuo in Nujiang Prefecture to the west of Dali, while Central Bai varieties were represented by Lanping Jinding (also in Nujiang), Jianchuan Diannan, and Heqing Jindun. Southern varieties include Dali Zhoucheng, Dali Qiliqiao, and Xiangyun Hedian. Allen (2004) placed Eryuan Xishan and Yunlong Baishi within the Central Bai group due to their high level of comprehension of other Central Bai varieties. However, Eryuan and Yunlong share the *Tone 1a innovation with other Southern varieties in contrast to Central Bai, so historically they belong to Southern Bai. See Figure 1 for a map of the Bai language area.

Allen also recorded personal narratives in seven of the nine locations (excluding Xiangyun and Qiliqiao) and used them to develop intelligibility tests known as Recorded Text Tests (RTT) (see Section 3.1). We correlate Allen's RTT scores with the Levenshtein distances based on his wordlists.

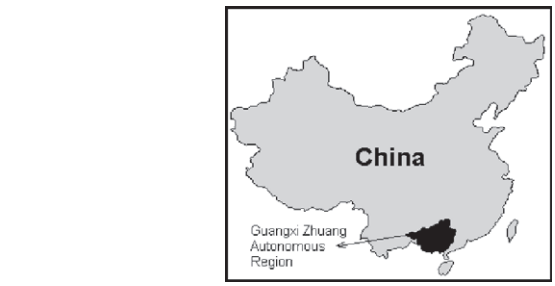


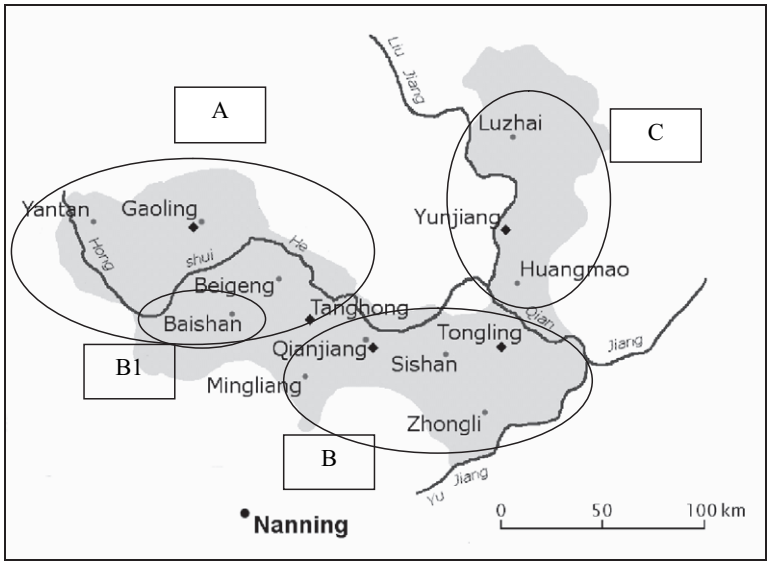
Source: Adapted from Allen, 2004; reprinted with permission

Figure 1. Bai language area.

2.2 Hongshuihe Zhuang

Hongshuihe (HSH) Zhuang, with a population of over 3 million, is spoken in Guangxi Zhuang Autonomous Region in southern China (National Statistics Department, 2003). HSH Zhuang is a Northern Tai language of the Tai-Kadai family. HSH Zhuang is a subgroup of the Zhuang ethnic group, but remains linguistically distinct from other Zhuang languages. See Figure 2 below for the location of Guangxi and HSH Zhuang language area.





Source: Adapted from Castro and Hansen (accepted); reprinted with permission

Figure 2. HSH Zhuang language area.

Very little dialect research has been done on HSH Zhuang varieties, but in 2006 Castro and Hansen (accepted) collected 511-item wordlists in 20 locations. Through analysis of the different diachronic developments in the dialect areas' tone systems, they identify three major dialect areas: A, B, and C, seen in Figure 2 below. Using Li's (1977) reconstruction of Proto-Tai tones, Castro and Hansen group the varieties together by tracking their modern reflexes for historic tone categories. Li reconstructs three Proto-Tai tone categories for syllables ending in a vowel or nasal (known as 'unchecked syllables'), which he labels

Table 1. Divergent diachronic developments in HSH Zhuang tone systems.

	Area A	Area B	Area B1	Area C
Varieties: / Tone category:	Yantan, Gaoling, Beigeng, Tanghong	Mingliang, Qianjiang, Sishan, Zhongli, Tongling	Baishan	Yunjiang, Luzhai, Huangmao
A1	53	35	35	53, in some places 53ʔ
A2	21, in some places 231	13	31	231
B1	33	55	55	45
B2	31	31	21	31ʔ
C1	453	33	45	53, in some places 453
C2	13, in some places 213	21	13	213

Source: adapted from Castro and Hansen (accepted).

A, B, and C. Voiced initials produced a lower pitch, which became phonemic tone when the voicing distinction was lost, producing six categories: A1, B1, and C1 for syllables with voiceless initials, and A2, B2, and C2 for syllables with originally voiced initials. Table 1 shows the divergent modern reflexes for each tone category in each area, as well as the varieties included in each area. For Areas A and B, the contour shapes are often the converse of each other: Area A's high falling tone in A1 corresponds to Area B's high rising, etc. Area C tones are distinct from the other two groups by the added feature of glottal closure at the end of the contour in tones A1 and B2. Castro and Hansen tentatively group Area B1 with Area B, but the tone correspondences do not match completely.

In each dialect area, Castro and Hansen selected one or two varieties that shared the most lexical items with other locations in their respective areas, for a total of five varieties: Gaoling and Tanghong for A, Qianjiang and Tongling for B, Yunjiang for C. In Figure 2 below, these sites are marked with black diamonds. They then used the personal narratives collected in those five varieties to conduct intelligibility testing (RTT methodology, see Section 3.1) in twelve locations: Yantan, Gaoling, Beigeng, and Baishan in A, Qianjiang, Mingliang, Sishan, Tongling, and Zhongli in B, and Yunjiang, Luzhai and Huangmao in C. See Figure 2 below for their locations. The twelve testing locations were selected in order to provide the widest geographic representation of HSH Zhuang.

Table 2. Cost of operations in Levenshtein distance.

Location	Transcription	Operation	Cost
Eryuan [k^{j}e^{35}]	k^{j}e_MRH		
	ke_MRH	delete j	1
	ke_HRH	substitute H for M	1
Jianchuan [ke^{55}]	ke_HEH	substitute E for R	1
		total cost	3
		normalised cost	.23

3. METHODOLOGY

To answer our first question about the relation between tone and intelligibility, we calculate the Levenshtein distances for tone separately from that of phonetic segments, correlate the two distances with intelligibility, then use multiple regression analysis to see which variable has the greater contribution to intelligibility. To answer our second question about the optimal representation of tone, we first calculate the Pearson correlation coefficient between intelligibility and the Levenshtein distances using each of the four representations of tone described below in Section 3.1.1. We then use Meng et al's (1992) test for correlated correlation coefficients to see if the highest coefficient is significantly higher than the others.

3.1 Levenshtein distance

Levenshtein distance, also known as edit distance, measures the difference between two strings. The Levenshtein distance between phonetic transcriptions of two cognates is the least number of insertions, deletions and substitutions needed to transform one cognate into another (Heeringa, 2004). One of the strengths of Levenshtein distance is that it always uses the string alignment that incurs the least cost of transformation. We applied Levenshtein distance to the Bai (500-item) and HSH Zhuang (511-item) wordlists using the free RuG-L04 software developed by Kleiweg (2004).

Table 2 calculates the Levenshtein distance between two Bai cognates for the gloss 'chicken,' pronounced [ke^{55}] in Jianchuan and [k^{j}e^{35}] in Eryuan. Tone is represented with tone onset, contour, and offset (see 3.1.1). Insertions, deletions, and substitutions are all weighted the same, with the cost of one. Differences are binary: either the sounds are the same, with no transformation cost, or they are different, with a cost of one. Heeringa et al (2006) found that using binary segment differences was equally valid with that of using gradual measures of distance between segments.

In order to prevent longer words from having undue weight in the calculation of average distance, a normalisation function was used in which the total cost is

Table 3. Sub-lexical variation in HSH Zhuang.

English	Proto-Tai	Liushui	Yantan	Gaoling	Mingliang
chin	khaaŋ A2	haːŋ^{231}	haːŋ^{23} kaːu^{44}	haːŋ^{23}pɛ^{433}	haːŋ^{31} ma^{31}

divided by the sum of length of the two strings. In cases where the two strings have no segment in common, the Levenshtein distance is proportional to the sum of the length of the two strings (Kleiweg, 2004).

When measuring phonetic distance, lexical variation is filtered out, and only cognates are accepted as data. However, Bai and HSH Zhuang are isolating languages with many di-syllabic words, in which one syllable may be cognate with other dialects, but the other syllable may not be. Table 3 shows HSH Zhuang words for 'chin', Proto-Tai *khaaŋ A2, in which the first syllable in each variety is the modern reflex of the Proto-Tai, but the second syllables are not cognate with each other. To filter out this 'sub-lexical' variation would place the Levenshtein distance at the morpheme level, not the lexical level, thus removing it even further from the context in which communication occurs. Therefore, we chose not to remove the non-cognate syllables in di-syllabic words. When two words had no overlapping syllables, we treated them as different lexemes.

3.1.1 Representing tone in Levenshtein distance

Phonologists often represent tone as autosegments of H(igh) and L(ow) attached to a tone-bearing unit (Duanmu, 1994; Zsiga and Nitisaroj, 2007). Contour tones like rising or falling are represented as a sequence of autosegments, LH for rising and HL for falling. Xu and Wang (2001), however, propose a Target Approximation (TA) model, in which targets can be static (register tones, e.g. H or L), or dynamic (contour tones, e.g. R for rising). We include both autosegment and target representations of tone in our evaluation.

Autosegments and target representations use salient perceptual cues for native speakers, but the listeners in the intelligibility tests described in section 3.2 are not native speakers of the language they are listening to, but rather speakers of related varieties. Burnham and Francis (1997) suggest that non-native speakers use multiple cues to identify tone, whereas native speakers have already fine-tuned their perception system down to the most salient features. Therefore, a more explicit way of representing tone is introduced, which includes the tonal onset (H for high, M for mid, L for low), contour shape (R for rising, F for falling), and tonal offset.

This moves the representation of tone to a more phonetic level, rather than the more abstract, phonemic level used in autosegmental and target representations. The phonetic representation of onset-contour-offset breaks tone down into three parts, making tone more specified than segments, whose differences are treated

Table 4. Comparison of the impact on Levenshtein distance of different tone representations.

Representation	Southern Bai	Central and Northern Bai	Levenshtein distance
onset-contour-offset	MRH	HEH	2
onset-contour	MR	HE	2
contour-offset	RH	EH	1
target	R	H	1
autosegments	_H	LH	1
Chao's pitch numbers	35	55	1

as binary, as seen in Table 2. Since the specified features of tone may be important perceptual cues to the listeners, the unequal treatment of tone and segment is acceptable for the purposes of this research.

We also include two variations of this more explicit representation: onset-contour and contour-offset. Both the autosegment and TA models agree that it is the offset, or target pitch, of the tone that is most salient to the production and perception of native speakers. However, if the listeners operate more as non-native speakers, they may also use tonal onset as a perceptual cue.

For East Asian tone languages, linguists usually use Chao's (1930) system of transcription. This notation treats tone as a sequence of pitch levels (5 for high, 1 for low), which may not be the optimal representation when modelling intelligibility. The disadvantage of this system is that it treats tone as a series of discrete pitch levels, leaving out explicit information about the contour shape that listeners may use in cross-dialectal comprehension. We include this representation in our evaluation for completeness.

Table 4 shows an example of the impact these various representations of tone have on Levenshtein distance. Southern Bai's mid-rising tone corresponds to Central and Northern Bai's high tone in syllables with Proto-Bai *voiceless unaspirated or *voiced sonorant initials in *Tone 1a. While onset-contour-offset and onset-contour representations count the cost as 2, all other representations count the cost as only 1. This example shows that different representations of tone have a substantial impact on how Levenshtein distance is computed.

3.2 Intelligibility tests

For both the Bai and HSH Zhuang surveys, intelligibility was tested with Recorded Text Tests (RTT), a method developed by Casad (1974) and further refined by Blair (1990). An RTT is a short personal narrative recorded in dialect A and played to a listener from dialect B. The dialect B listener answers content

questions about the narrative, and the percentage of correct answers is interpreted as their comprehension level of dialect A.

Allen chose seven sites to record personal narratives in, three sites each from Central and Southern Bai and one from Northern Bai (see Figure 1 for site locations). Castro and Hansen chose five sites, two from dialect area A, two from B, and one from C (see Figure 2 for site locations). Sites were selected to represent speaker population centers, as well as the geographic and linguistic spread of the language. At each site, a native speaker was recorded telling a personal narrative two to five minutes in length, and then the story was translated into the national language (Chinese). The researcher developed 12–20 content questions and asked five to ten local speakers to listen to the narratives and answer the questions. Any question not correctly answered by a local speaker was deemed to be defective, i.e. an irrelevant question that arose from a faulty translation of the text. Defective questions were not included in later testing.

After pilot-testing the narrative recordings, the researcher selected test sites. Allen returned to the same sites as the recording locations, except Luobenzhuo in Northern Bai. Shared vocabulary of less than 60 per cent between Luobenzhuo and all other varieties was deemed a significant enough difference as to make intelligibility unlikely. Allen tested the other six sites on the Luobenzhuo recording with results indicating very low intelligibility. In total, seven tests were conducted in six testing locations (see Figure 1). Castro and Hansen tested five stories at 12 testing sites (see Figure 2); though they did not test every recording at each testing site, they did obtain results for 44 pairs of varieties.

At each testing site, the researcher asked approximately eight to twelve native speakers to listen to the recordings individually using headphones and answer content questions about the text. The listeners were selected to ensure a balance between male and female listeners, between older and younger speakers, and between those with more and less education. Listeners were native speakers who were born and raised in the village, whose parents were both native speakers from the village, and who had minimal contact with speakers from other varieties. Each recording was played twice; the second time, content questions were asked after each section using either the listener's variety or Chinese. If listeners responded with an incorrect answer, the relevant section was played again. Most tests consisted of ten questions, each question being asked immediately after the section that contained the answer. Percentage of correct answers out of the total number of questions constituted the intelligibility test score. Correct answers got one point, while answers that were considered partially correct got half a point.

Tests were individually and orally administered, which substantially increased the time needed for each participant. However, playing the stories for larger groups and asking participants to write down their answers was not viable, since a written test assumes a rate of literacy in Chinese that is unlikely among

Table 5. Pearson's correlation and explained variance of tone and segment levels with respect to intelligibility.

Language	Variable	Correlation (r)	Explained variance ($r^2 \times 100$, %)	Significance
Bai	Tone	−0.75	56	< 0.0001
	Segments	−0.71	51	< 0.0001
HSH Zhuang	Tone	−0.66	43	< 0.0001
	Segments	−0.68	46	< 0.0001

Table 6. Results of multiple linear regression analysis, in which intelligibility is the dependent variable and tone and segments are the independent variables.

Language	Variable	*t*-value	Significance
Bai	Tone	−2.040	0.049
	Segments	−0.604	0.550
HSH Zhuang	Tone	−2.255	0.030
	Segments	−2.934	0.005

ethno-linguistic communities in rural China. Allen had an average sample of ten participants per site, with a total of 419 participants. Castro and Hansen had an average of eight participants, with a total of 372 participants.

4. RESULTS

4.1 Multiple regression analysis for tone and segments

The segmental and suprasegmental distances are each correlated separately with intelligibility test scores for Bai and HSH Zhuang. The correlations are given in Table 5. Tone is represented by onset-contour, since that representation shows a stronger correlation with intelligibility (see Section 4.2 below). For both languages, tone and segments both have a strong, significant correlation with intelligibility.

Multiple regression analysis reveals the relative contribution each level makes to intelligibility, seen in Table 6. For Bai, tone is the main predictor of intelligibility, even more important than the segmental level, which surprisingly appears as an insignificant variable. In contrast, for HSH Zhuang, both tone and segment levels are significant variables in relation to intelligibility. For both languages, tone makes a significant contribution to intelligibility, but in Bai the results suggest that tone is more important than segments. These results suggest

Table 7. Comparison of tone representations' correlation with intelligibility for Bai and HSH Zhuang.

Bai	Correlation (r)	Explained variance ($r^2 \times 100$, %)	Significantly higher than others?
onset-contour	−0.75	56	yes, $p < .001$
onset-contour-offset	−0.75	56	yes, $p < .001$
contour-offset	−0.75	56	yes, $p < .01$
target	−0.74	54	no
autosegments	−0.74	54	no
Chao's pitch numbers	−0.72	52	no
HSH Zhuang	**Correlation (r)**	**Explained variance ($r^2 \times 100$, %)**	**Significantly higher than others?**
onset-contour	−0.72	52	yes, $p < .001$
onset-contour-offset	−0.71	50	yes, $p < .05$
Chao's pitch numbers	−0.68	46	no
target	−0.65	42	no
autosegments	−0.63	40	no
contour-offset	−0.62	38	no

that the contribution of tone to intelligibility is significant across East Asian tone languages, and for some languages tone may actually be the main predictor of intelligibility.

4.2 Comparison of different representations of tone

Table 7 shows the Pearson correlation coefficients (r) for the various representations of tone described in Section 3.1.1 and intelligibility in Bai and HSH Zhuang, as well as the percentage of variance explained (R^2). All correlation coefficients are highly significant ($p < .0001$).

For both Bai and Zhuang, the correlations for onset-contour and onset-contour-offset are significantly higher than other representations. In both languages, onset-contour is higher at a significance level of $p < .001$, while onset-contour-offset is higher at a significance level of $p < .001$ for Bai and of $p < .05$ for Zhuang. These results suggest that representing tone with onset-contour and onset-contour-offset better approximates intelligibility than other representations.

In Bai, the correlation for contour-offset was also stronger than others, at a significance level of $p < .05$. When onset-contour, onset-contour-offset, and contour-offset were compared to each other, without considering target, Chao,

or autosegments, none of them was significantly higher. Likewise, for Zhuang, neither onset-contour nor onset-contour-offset was higher than the other.

5. CONCLUSIONS

We applied Levenshtein distance to Bai dialect data from Allen (2004) and HSH Zhuang dialect data from Castro and Hansen (accepted). We correlated Levenshtein distance with intelligibility test scores and found significant inverse correlations in both languages. These findings suggest that Levenshtein distance is a useful tool for dialectologists working with East Asian tone languages, and that Levenshtein distance provides a good model for predicting intelligibility in these languages.

Gooskens and Heeringa (2006) found that the relative contribution of prosody in Norwegian to perceived distance was negligible, but they did not investigate the relative contribution to intelligibility. Tang and van Heuven (2007) assert that in a full-fledged tonal language, like Chinese, the contribution of tone to perceived intelligibility and perceived distance is greater than that found for Norwegian, which only has two tonemes. This research affirms Tang and van Heuven's assertion; we find a strong correlation between tonal information and intelligibility, with tone alone able to explain 43 percent of the variance in HSH Zhuang and 56 percent in Bai. This suggests that, in tonal languages, differences in tone have a significant impact on intelligibility. It would be interesting to compare the relative contributions of consonants, vowels, and tones to intelligibility to see if tone is the single most salient factor in cross-dialectal comprehension in tonal languages.

Finally, we used different representations of tone in Levenshtein distance to see which would show the strongest correlation to intelligibility. For both languages, phonetically explicit representations of tone that included information on tonal onset and contour were superior to others. Representations at a more unspecified, phonemic level such as autosegments and targets showed a weaker correlation with intelligibility. This suggests that cross-dialectal listeners operate as non-native speakers who use a variety of perceptual cues, rather than as native speakers who have narrowed in on a subset of salient features.

For Zhuang, onset and contour were essential ingredients for an optimal approximation of intelligibility, while contour-offset was shown to be less than optimal. But for Bai, contour-offset proved superior to autosegment, target, and Chao's pitch numbers. Thus, at this stage we cannot confirm which is more perceptually salient to cross-dialectal listeners, the onset or the offset. However, given that both Bai and Zhuang show onset-contour as superior, whereas contour-offset is shown only to be optimal in Bai, preference may be given to onset-contour representations.

One explanation for why onset-contour is optimal can be found by reviewing the impact of different tone representations illustrated in Table 4 above (see Section 3.1.1). The distinguishing innovation for Southern Bai is the mid rising (35) tone corresponding to Central and Northern Bai's high level (55) tone. Both onset-contour-offset and onset-contour weighs the distance between 35 and 55 at least twice the amount of the other representations. Even though these two tones end up at the same pitch level, they begin at different levels and have different contour shapes, making them difficult for cross-dialectal listeners to assimilate into their own system. Thus, the greater the weight assigned to this kind of difference, the better the inverse correlation with intelligibility.

This research has relevance for the dialectometry of other East Asian tone languages, many of which have not yet been the object of serious study. Using an optimal representation of tone in Levenshtein distance may result in dialect groupings that are more coherent with the perceptions of listeners, and therefore from a more emic perspective. Our findings also have implications for tone perception research, which has mainly focused on native speakers of major languages such as Thai and Chinese (Gandour et al., 2000; Burnham and Francis, 1997; Zsiga and Nitisaroj, 2007). Since cross-dialectal listeners are somewhat in between native and non-native, perceptual studies of such listeners may reveal that they use a wider range of perceptual cues than native speakers, as suggested by this research.

ACKNOWLEDGEMENTS

We are grateful to David Bradley and Eric Jackson for comments on this work and to the participants at the Methods in Dialectology XIII workshop for their constructive scrutiny. Three anonymous reviewers also gave insightful and helpful comments.

REFERENCES

B. Allen (2004), *Bai dialect survey*, Translated by Zhang Xia. Yunnan Minority Language Commission and SIL International Joint Publication Series (Kunming).

F. Blair (1990), *Survey on a Shoestring*, SIL and UTA Publications in Linguistics 96 (Dallas).

D. Bradley (2007), 'East and South East Asia', in R. E. Asher and C. Moseley, eds, *Atlas of the World's Languages* (London).

D. Burnham and E. Francis (1997), 'Role of linguistic experience in the perception of Thai tones', in A. S. Abramson, ed., *Southeast Asian linguistic studies in honor of Vichin Panupong* (Bangkok).

E. H. Casad (1974), *Dialect intelligibility testing* (Dallas).

A. Castro and B. Hansen (accepted), 'Hongshuihe Zhuang dialect intelligibility survey*', SIL Electronic Survey Reports*.

Y. Chao (1930), 'A system of tone letters', *Le Maitre Phonetique*, 30, 24–27.

S. Duanmu (1994), 'Against contour tone units', *Linguistic Inquiry,* 25.4, 555–608.

J. Gandour, D. Wong, L. Hsieh, B. Wienzapfel, D. Van Lancker, and G. D. Hutchins (2000), 'A crosslinguistic PET study of tone perception', *Journal of Cognitive Neuroscience*, 12, (1), 207–222.

C. Gooskens (2006), 'Linguistic and extra-linguistic predictors of inter-Scandinavian intelligibility', in J. van de Weijer, and B. Los, eds, *Linguistics in the Netherlands 2006* (Amsterdam), 101–113.

C. Gooskens and W. Heeringa (2004), 'Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data', *Language Variation and Change*, 16, (03), 189–207.

C. Gooskens and W. Heeringa (2006), 'The relative contribution of pronunciational, lexical, and prosodic differences to the perceived distances between Norwegian dialects', *Literary and Linguistic Computing*, 21, (4), 477–492.

W. Heeringa (2004), 'Measuring pronunciation differences with Levenshtein distance' (PhD thesis, University of Groningen).

W. Heeringa, P. Kleiweg, C. Gooskens, and J. Nerbonne (2006), 'Evaluation of string distance algorithms for dialectology', in John Nerbonne and Erhard Hinrichs, eds, *Linguistic Distances Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics* (Sydney), 51–62.

B. Kessler (1995), 'Computational dialectology in Irish Gaelic', *Proceedings of the European ACL* (Dublin), 60–67.

P. Kleiweg (2004), *RuG/L04, software for dialectometrics and cartography* [computer program]. [2004] < URL: http://www.let.rug.nl/∼kleiweg/indexs.html. > [24 May 2008].

F. Li (1977), *A Handbook of Comparative Tai* (Honolulu).

J. A. Matisoff (2003), *Handbook of Proto-Tibeto-Burman: System and Philosophy of Sino-Tibetan Reconstruction*. UC Publications in Linguistics, 135 (Berkeley).

X. L. Meng, R. Rosenthal, and D. B. Rubin (1992), 'Comparing correlated correlation coefficients', *Psychological Bulletin*, 111, (1), 172–175.

National Statistics Department (2003), *Nian renkou pucha Zhongguo minzu renkou ziliao [Information on China's minority population from the 2000 census]* (Beijing).

C. Tang and V. J. van Heuven. (2007), 'Mutual inteligibility and similarity of Chinese dialects: Predicting judgments from objective measures', in B. Los and M. van Koppen, eds, *Linguistics in the Netherlands* (Amsterdam), 223–234.

F. Wang (2005), 'On the genetic position of the Bai language', *Cahiers de Linguistique – Asie Orientale*, 34, 101–127.

F. Wang (2006), *Comparison of languages in contact: the distillation method and the case of Bai*. Language and Linguistics Monograph Series B: Frontiers in Linguistics III (Taipei).

L. Xu and Y. Zhao (1984), *Baiyu Jianzhi [Description of the Bai Language]* (Beijing).

Y. Xu and Q. E. Wang (2001), 'Pitch targets and their realization: Evidence from Chinese', *Speech Communication*, 33, 319–337.

C. Yang (accepted), 'Nisu dialect geography', *SIL Electronic Survey Reports*.

Yunnan Provincial Statistics Department (2004), *Yunnan Statistical Yearbook* (Beijing).

E. Zsiga and R. Nitisaroj (2007), 'Tone features, tone perception, and peak alignment in Thai', *Language and Speech*, 50, (3), 343–383.