2023. *Proc Ling Soc Amer* 8(1). 5465. https://doi.org/10.3765/plsa.v8i1.5465.

Decision trees, entropy, and the contrastive feature hierarchy

Jane Chandlee*

Abstract. Dresher (2009) argues that language-particular hierarchies of features are the best way to identify contrastive features in a phonological inventory. While not universal, this ordering of features is also not fully unconstrained. But what limits the space of possible feature orders remains an open question. This paper demonstrates how the concept of entropy establishes a partial ordering of features that both allows for but also constrains language-particular variation. Specifically, a decision tree machine learning algorithm is employed to dynamically impose structure on the hypothesis space of possible feature orders.

Keywords. phonological features; contrastive hierarchy; entropy; machine learning

1. Introduction. The Contrastive Hierarchy is a proposal for specifying the contrastive features of a language's segment inventory by recursively dividing the sounds based on a prescribed feature order (Dresher 2009). The accompanying Successive Division Algorithm (SDA) creates a tree structure in which the leaves are the sounds and non-leaf nodes contain subsets of the inventory that require further specification by additional features. Features not included on the path to a leaf are either left unspecified or else can be specified but considered redundant.

For example, based on a pattern of ATR harmony, Mackenzie & Dresher (2003) propose the feature order in (1) for the vowel inventory of Nez Perce (Sahaptian; Idaho, Washington, Oregon):

(1) low > round > ATR

Following Hall & Hall (1980), they assume the inventory includes {æ, ɑ, i, u, ɔ, ɛ}, where /ɛ/ is an abstract −ATR vowel that merges with [i] on the surface. The order in 1 then gives the tree shown in Figure 1. This tree is essentially a decision tree, a connection made more explicit by Cherry et al. (1953), who conceive of phoneme identification as a series of yes-no questions, one per feature.

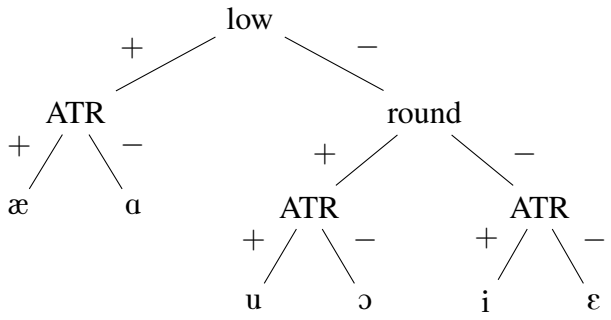

Figure 1. Contrastive specifications of Nez Perce vowels proposed by Mackenzie & Dresher (2003)

Feature orders are language-particular and are identified based on patterns of distribution and alternations (i.e., which features are phonologically active). This means that languages with

* Haverford College (jchandlee@haverford.edu)

the same segment inventory may still differ with respect to the ordering of features and therefore the resulting specifications. At the same time, possible feature orders are not believed to be fully unconstrained. Writes Dresher (2009:168): 'The limits of this variation [in feature orders] remain to be determined...That there are limits is suggested by the fact that certain feature orders produce unnatural-looking inventories'. This paper demonstrate how the concept of entropy can provide a compromise between a fixed universal ordering and completely free ordering. In particular, a modified version of a decision tree learning algorithm is used to establish a partial ordering of features that constrains language-particular variation.

2. Proposal. The machine learning algorithm ID3 (Iterative Dichotomiser 3; Quinlan 1986) learns decision trees by recursively selecting the feature with the largest *information gain*, defined as the greatest reduction in system entropy. Informally, entropy can be thought of as the degree of uncertainty in identifying the correct phoneme based on the information (i.e., feature specifications) available so far. Formally, it is calculated using the formula in (2),

(2) $\sum_{i=1}^{N} -p_i \log_2 p_i$

where p_i is the proportion of decisions/phonemes in a given category and N is the total number of categories. At the start (i.e., the root of the tree), with no features specified, entropy is at its greatest point. Each segment is its own category and so the entropy is N (- 1/N $\log_2$ 1/N), where N is the number of segments in the inventory.

Adding features reduces entropy by grouping the segments into natural classes and therefore reducing the number of categories. ID3's feature selection is greedy in that it selects the feature that reduces entropy by the greatest amount (i.e., by providing the most information). In practice this will be the feature that divides the current set of segments most evenly.

What happens when multiple features are equally informative? Applying a universal order would have the undesirable consequence that languages with the same inventory will have the same contrastive features, contrary to fact. Instead, the proposal is that these clusters of equally informative features are exactly where language-particular variation is possible. Put another way: if one feature has the greatest information gain, it must be selected, but when multiple features tie for information gain, the language can choose freely among them. The result is a default partial order based on entropy (which is itself based on the structure of the inventory), with language-particular variation only possible when information gain is indecisive.

Importantly, ID3 as employed here is not learning the actual feature order for a particular language. As noted above, that order is determined using cues from distributions and alternations, information that the algorithm doesn't have access to. Rather, the proposed contribution of ID3 is the way its method of feature selection limits the space of possible feature orders. The following demonstrations will make this idea more clear, culminating in a hypothesis for possible orders.

3. Test case 1: Nez Perce. To see how this works we'll return to Mackenzie & Dresher's (2003) analysis of Nez Perce mentioned above (see also Aoki 1966; Baković 2000). Again the proposed order is low > round > ATR. Given the feature specifications shown in Table 1, the run of ID3 proceeds as depicted in Figure 2. Selected features are shown in bold; equally informative but not selected features are in italics.

Initially, the feature ATR is selected automatically as the most informative feature for the entire inventory. For the +ATR vowels {æ, i, u}, all of the features round, high, low, and back

	−back		+back	
	+ATR	−ATR	+ATR	−ATR
+hi,−lo	i		u	
−hi,−lo		ɛ		ɔ
−hi, +lo			æ	ɑ

Table 1. Feature specifications for Nez Perce vowels (gray shading = +round)

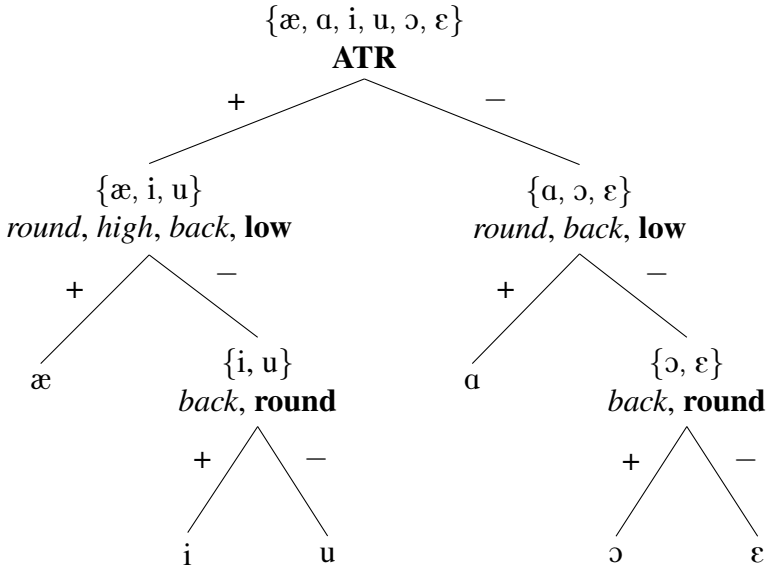

Figure 2. Run of ID3 on Table 1

are equally informative. Among these, low is the highest feature in the proposed order, and so it is chosen. Likewise for the remaining internal nodes of the tree: among the features determined to be equally informative, the highest one according to the proposed order is chosen. Note that while the structure of this tree differs from the one in Mackenzie & Dresher (2003) (i.e., the one in Figure 1), the resulting feature specifications are the same.[1]

The consequences of this result are the following. First, it confirms that ID3 modified with this choice procedure can generate the desired feature specifications for this language. And second, we have a prediction for how other languages with the same (or an equivalently-structured) inventory could vary in their feature order. According to Figure 2, after ATR, any of {round, high, low, back} could be ordered. Given that this set includes all of the remaining features under consideration, this prediction is not particularly insightful. But more can be gleaned from the additional demonstrations on larger inventories to follow.

4. Test case 2: Bumo Izon. The second test case is based on Bumo Izon (Niger-Congo; Nigeria), which has a co-occurrence restriction on implosives and voiced plosives (Efere 2001; Hansson 2001; Harry 2004). To account for this pattern, Mackenzie (2009) proposes the constraint *[αc.g.][$-\alpha$c.g.]$_{\text{Root}}$ and the feature order in (3).

[1] Indeed Mackenzie & Dresher (2003) observe that it doesn't really matter where ATR is situated in the order, since the inventory is symmetric with respect to this feature.

(3) labial > dorsal > voice > c.g.

The consequence of this order is that the segments that do not participate in the restriction (velars and labiovelars) are not specified for c.g. The relevant portion of the consonant inventory is given in Table 2.

	labial	dorsal	coronal
+c.g.	ɓ		ɗ
	gɓ		
−c.g.	p b	k g	t d
	kp		

Table 2. Subset of Bumo Izon consonant inventory

ID3's run on this inventory is illustrated in Figure 3. Again the choice among features with the same information gain is made following (3). The resulting tree is in fact exactly the one proposed by Mackenzie (2009:31).

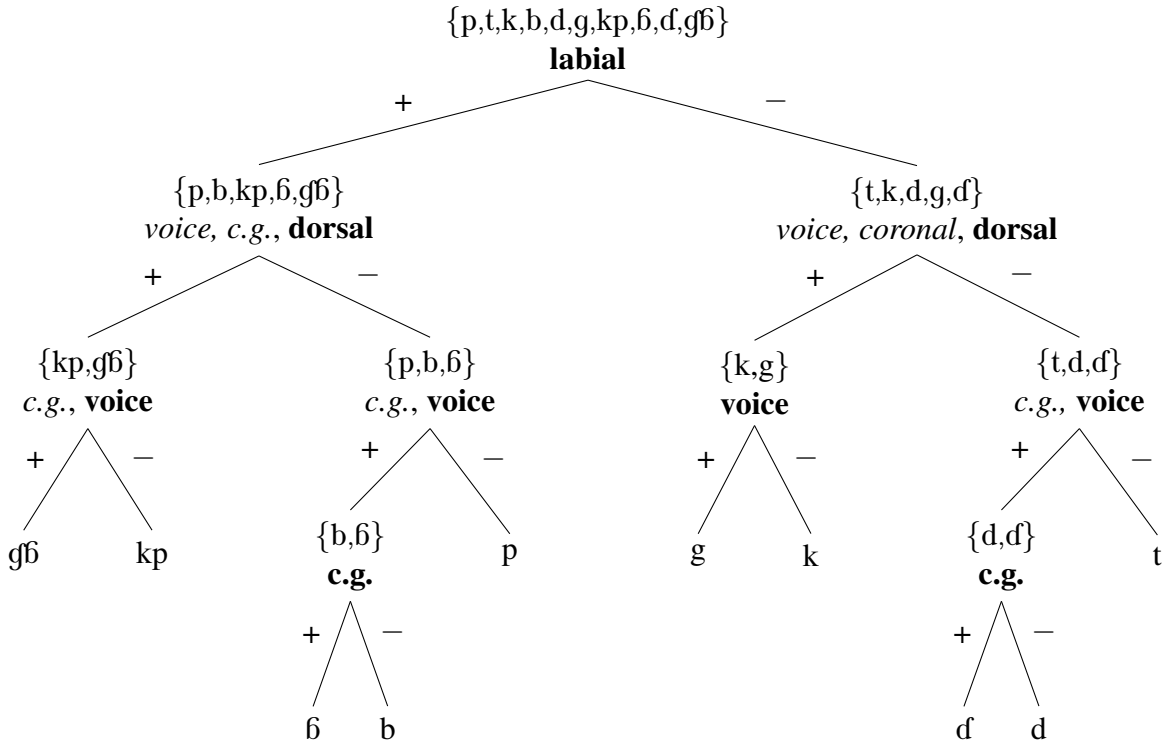

Figure 3. Run of ID3 on Table 2

5. Test case 3: Tzutujil. The third and final test case is based on another consonant co-occurrence restriction, this one from Tzutujil (Mayan; Guatemala). Ejectives cannot co-occur unless identical, nor can homorganic ejective-plain stop pairs (MacEachern 1999). Implosives, however, can occur freely with other segments. Mackenzie (2009) accounts for these patterns with the feature order in (4), which prevents implosives from being specified for the feature c.g.

(4) labial > dorsal > coronal > voice > c.g.

Mackenzie demonstrates this feature order using the subset of the inventory shown in Table 3.

	labial	coronal	dorsal	glottal
+c.g.	ɓ	ɗ	k’	ʔ
−c.g.	p	t	k	

Table 3. Subset of Tzutujil consonant inventory

Given the same subset of the inventory, ID3 gets the specifications exactly wrong! As shown in Figure 4, c.g. is selected automatically as the initial, most informative feature. As a result, all segments—including the implosives—are specified for c.g.

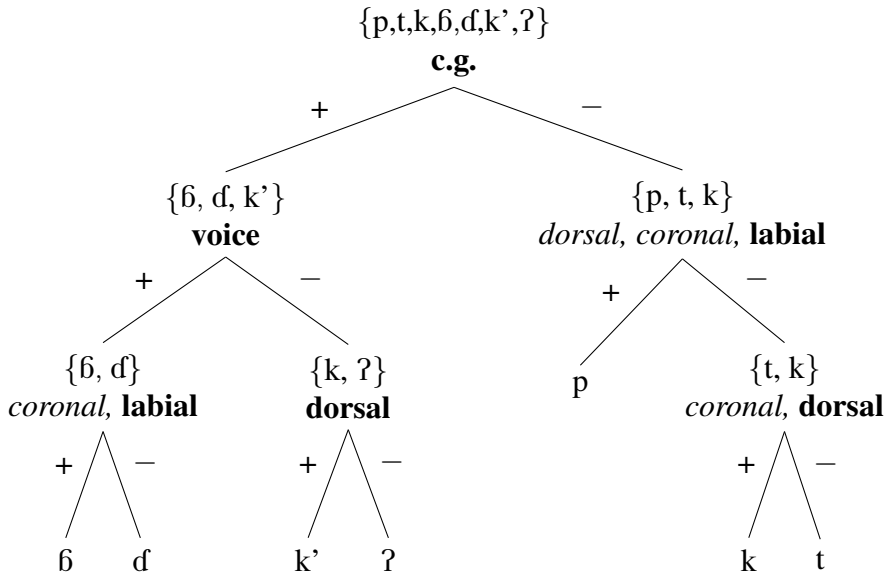

Figure 4. Run of ID3 on Table 3

However, things get more interesting if we instead start from the larger inventory listed in Table 4, which includes all obstruents except for affricates (Dayley 1985). This larger set of course requires additional features. Following common practices, the feature continuant is used to distinguish stops from fricatives. To distinguish alveolars and post-alveolars, the secondary feature anterior is used and only specified for segments that are +coronal. To distinguish velars and uvulars, the secondary feature high is used and only specified for segments that are +dorsal. Glottals are represented as [−labial, −coronal, −dorsal].

The run of ID3 on this expanded inventory is illustrated in Figure 5. This time, in the resulting tree the implosives are not specified for c.g., as desired. Only the pairs that differ only in this feature are specified for it: the velars {k, k’} as in Mackenzie (2009:98)’s tree, and now also the uvulars {q, q’}.

The specification of place features, however, does not align with common conventions. In particular, the features high and anterior are used *instead* of dorsal and coronal, rather than as

	labial	coronal ant		dorsal hi		glottal
+c.g.	ɓ	ɗ		k’	q’	ʔ
−c.g.	p	t		k	q	
+cont		s	ʃ		χ	h

Table 4. Tzutujil obstruents (minus affricates)

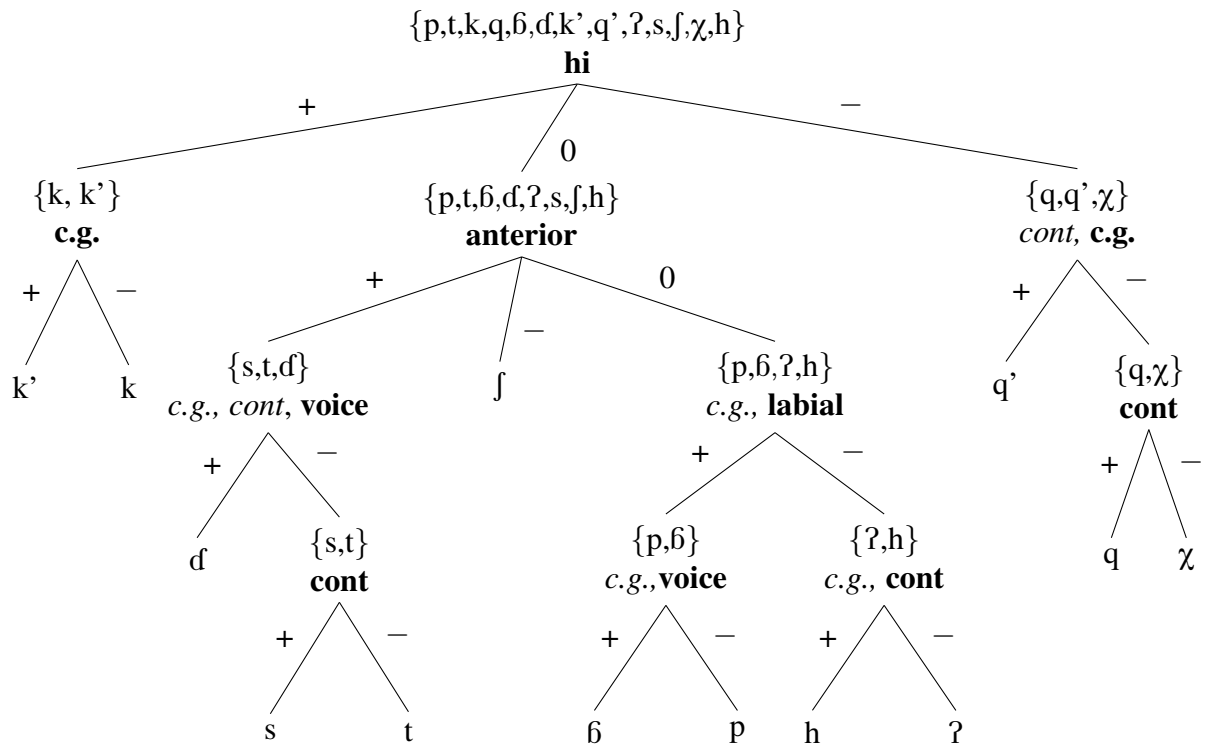

Figure 5. Run of ID3 on Table 4

secondary features. This is not surprising, as the current version of ID3 treats all features equivalently. These secondary features are simply treated as ternary primary features with the values $\{+, -, 0\}$. This makes them particularly attractive, since the three-way division of sounds results in a greater reduction in entropy. There are at least two options for how to address this. One is to further modify the algorithm so that when it selects a secondary feature, it must first select the corresponding primary feature. Another option is to instead modify the resulting tree to include the primary features. Determining which of these routes is preferable requires further testing on more cases in which place features play a crucial role in the feature ordering.

6. Additional tests. In addition to the three cases presented in the previous sections, ID3 was tested on the feature orders proposed for the vowel inventory of Classical Manchu (Zhang 1996; Dresher 2009) and the consonant inventories of Kalabari Ijo, Dholuo, Chaha, Anywa, Hausa, and Aymara, all analyzed by Mackenzie (2009).

In six of these cases, ID3 generated the exact same tree as in the original analysis, in one case it generated a tree that differed in structure but resulted in the same feature specifications

(that was Nez Perce), and in three cases the results differed in some respects but maintained the crucial specifications argued to capture the relevant co-occurrence constraint (for example the case of Tzutujil discussed above). These results are summarized in Table 5.

Language	Exact	Equivalent	Equal where it matters
Nez Perce		✓	
Classical Manchu	✓		
Bumo Izon	✓		
Kalabari Ijo	✓		
Dholuo	✓		
Chaha			✓
Anywa	✓		
Hausa			✓
Tzutujil			✓
Aymara	✓		

Table 5. Summary of results from all test cases

7. Discussion. As the above demonstrations have shown, ID3's use of information gain as a feature selection criterion constrains possible feature orders and leads to the following hypothesis for the limits on language-particular variation:

(5) A less informative feature will not need to be ordered above a more informative one.

More extensive testing of this hypothesis is needed, in particular with larger inventories. The example of Tzutujil above shows how the practice of using subsets of inventories for expository convenience is not always innocuous. The same example also highlighted a consequence of using a feature set with secondary features. More systematic testing with different feature sets could also raise interesting questions about this and other representational assumptions.

As noted earlier, ID3 does not learn the actual feature order for a given language; rather it uses that order to select among equally informative features.[2] Nonetheless, there is potential for it to serve as the foundation for a learner that does identify the feature order using a similar principle. In addition, its input is the segment inventory fully specified for all available features. Adapting it to a model in which features are instead emergent—see Mielke (2008)—would be an interesting way to bring it more in line with an actual model of acquisition.

8. Conclusions. If the hypothesis in (5) is correct, then ID3 serves as a valuable analytical tool for verifying the well-formedness of a proposed feature order. In addition, it provides a principled means of situating features into the hierarchy in those cases where language-internal cues don't dictate an exact position.

[2] See also Shwayder (2009) for an algorithm that aims to do both using the added input of which sets of segments contrast. Though it doesn't use entropy or information gain, this learner similarly selects features that most closely divide sets of segments in half. In the case of ties, a proposed universal hierarchy of features is consulted, followed by possible reranking based on the given knowledge about contrast.

More broadly, the results presented here suggest that the identification of distinctive features is another promising application of entropy and information theory, building on significant previous work on a variety of phonological learning problems (e.g., Goldwater & Johnson 2003; Goldsmith & Riggle 2012; Hume & Mailhot 2013).

References

Aoki, Haruo. 1966. Nez Perce vowel harmony and Proto-Sahaptian vowels. *Language* 42. 759–767. https://doi.org/10.2307/411831.

Baković, Eric. 2000. *Harmony, dominance and control*. New Brunswick: Rutgers University dissertation.

Cherry, E. Colin, Morris Halle & Roman Jakobson. 1953. Toward the logical description of languages in their phonemic aspect. *Language* 29(1). 34–46.

Dayley, Jon P. 1985. *Tzutujil grammar*. Berkeley: University of California Press.

Dresher, B. Elan. 2009. *The contrastive hierarchy in phonology*. Cambridge: Cambridge University Press.

Efere, Emmanuel. 2001. The pitch system of the Bumo dialect of Izon. In Suzanne Gessner, Sunyoung Oh & Kayono Shiobara (eds.), *African languages and linguistics*, 115–259. Vancouver: UBC Working Papers in Linguistics.

Goldsmith, John & Jason Riggle. 2012. Information theoretic approaches to phonological structure: the case of Finnish vowel harmony. *Natural Language & Linguistic Theory* 30. 859–896. https://doi.org/10.1007/s11049-012-9169-1.

Goldwater, Sharon & Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In Jennifer Spenader, Anders Eriksson & Östen Dahl (eds.), *Proceedings of the Workshop on Variation within Optimality Theory*, 111– 120. Stockholm: Stockholm University.

Hall, Beatrice L. & R. M. R. Hall. 1980. Nez Perce vowel harmony: An Africanist explanation and some theoretical consequences. In Robert M. Vago (ed.), *Issues in vowel harmony*, 201–236. Amsterdam: John Benjamins.

Hansson, Gunnar. 2001. *Theoretical and typological issues in consonant harmony*. Berkeley: University of California dissertation.

Harry, Otelemate. 2004. *Aspects of the tonal system of Kalabari-Ijo*. Stanford: CSLI.

Hume, Elizabeth & Frédéric Mailhot. 2013. The role of entropy and surprisal in phonologization and language change. In Alan C. L. Yu (ed.), *Origins of sound patterns: Approaches to phonologization*, 29–47. Oxford: Oxford University Press.

MacEachern, Margaret. 1999. *Laryngeal cooccurrence restrictions*. New York: Garland.

Mackenzie, Sara & B. Elan Dresher. 2003. Contrast and phonological activity in the Nez Perce vowel system. *Berkeley Linguistics Society (BLS)* 29(1). 283–291. https://doi.org/10.3765/bls.v29i1.979.

Mackenzie, Sara. 2009. *Contrast and similarity in consonant harmony processes*. Toronto: University of Toronto dissertation.

Mielke, Jeff. 2008. *The emergence of distinctive features*. Oxford: Oxford University Press.

Quinlan, J. Ross. 1986. Induction of decision trees. *Machine Learning* 1(1). 81–106.

Shwayder, Kobey. 2009. *The best binary split algorithm: A deterministic method for dividing vowel inventories into contrastive distinctive features*. Waltham, MA: Brandeis University MA thesis.

Zhang, Xi. 1996. *Vowel systems of the Manchu-Tungus languages of China*. Toronto: University of Toronto dissertation.