Investigating variation in written forms of Nahuatl using character-based language models

Robert Pugh and **Francis M. Tyers**
Department of Linguistics, Indiana University, Bloomington, IN
{pughrob,ftyers}@iu.edu

Abstract

We describe experiments with character-based language modeling for written variants of Nahuatl. Using a standard LSTM model and publicly available Bible translations, we explore how character language models can be applied to the tasks of estimating mutual intelligibility, identifying genetic similarity, and distinguishing written variants. We demonstrate that these simple language models are able to capture similarities and differences that have been described in the linguistic literature.[1]

1 Introduction

The diversity of language variants[2] in a linguistic continuum presents an interesting challenge to the development of language technology. For marginalized and endangered languages, the general lack of resources in the language as a whole exacerbates this challenge.

Character-level features have been shown to be effective for a wide range of textual NLP tasks, including language identification (Dunning, 1994; Veena et al., 2018), native language detection (Kulmizev et al., 2017), and machine translation (Lee et al., 2017; Chen et al., 2018). Furthermore, they offer the advantage of requiring little-to-no preprocessing or linguistic engineering (e.g. word tokenization, morphological segmentation, etc.) other than possibly orthographic normalization[3].

In this paper we investigate the usefulness of character language models in addressing questions about variation within a linguistic continuum. Specifically, we examine the extent to which these simple surface-level features of written language correspond to structural phonological and grammatical differences between different variants of Nahuatl. We examine three tasks: variant identification, linguistic sub-classification/genetic similarity, and the prediction of mutual intelligibility.

2 Background

In this section we give some background on the language, language modeling, and some relevant related work.

2.1 Nahuatl

Nahuatl is a polysynthetic, agglutinating Uto-Aztecan language continuum spoken throughout Mexico and Mesoamerica. The Mexican Government's *Instituto Nacional de Lenguas Indígenas* (INALI) recognises 30 distinct variants (INALI, 2009). These variants have highly-variable levels of intelligibility between them, and linguistic similarity and mutual intelligibility is not always correlated with geographic distance. Furthermore, the recognition and treatment of Nahuatl's linguistic diversity has far-reaching impacts on language activism and revitalization projects (Pharao Hansen, 2013).

Nahuatl variants can differ along lexical (*totoltetl* vs. *teksistli* 'egg'), phonological (common isoglosses include *t-tl-l* and *e-i*), and morphological (e.g. the presence or absence of word-initial *o-* for past tense verbs) dimensions, and orthography can vary within and across variants. Table 1 gives an example of these types of variation.

Computational modeling of Nahuatl variants is useful for many language technology applications. Automatic variant detection may be useful for grouping and categorizing texts in a large corpus such as the Axolotl corpus (Gutierrez-Vasques et al., 2016), where the provenance of the texts is not always known. For automated dialogue systems, variant modeling can be used to assess the degree to which a generated turn will be understood by a

[1] Code is available at https://github.com/Lguyogiro/nahuatl-variant-charlms-americasnlp

[2] We use the term *variants* to refer to instances of any kind of intra-language variation, including variation based on region (dialect), culture or ethnicity (ethnolect) etc. These may or may not be considered the same language or separate languages.

[3] Subword tokenization methods, such as Byte-Pair Encoding, also share this property. We leave the investigation of unsupervised subword tokenization for written Nahuatl for future experiments.

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 21–27
June 11, 2021. ©2021 Association for Computational Linguistics

user. Finally, for applications that generate text on a user's behalf, such as predictive keyboards and spell-checking systems, it is vital to maintain consistency in both language variant and orthographic norms.

2.2 Language Modeling

Language models are probability distributions over sequences of vocabulary items with parameters learned from data. They are ubiquitous in Natural Language Processing in areas including machine translation, automatic speech recognition, and spelling correction, among others. Traditional n-gram language models estimate the conditional probability of each vocabulary item given contexts of preceding vocabulary items based on their frequency in the data. Neural language models represent each vocabulary item as a distributed feature vector, and learn the joint probability function of the sequence of feature vectors and the feature vectors themselves simultaneously (Bengio et al., 2000). We use the latter in the experiments presented in this paper.

$$\mathrm{PP} = e^{-\frac{1}{N}\sum_{i=1}^{N} \ln p(x_i)} \qquad (1)$$

A common metric for evaluating the performance of a language model is perplexity (1), or how "surprised" the model is when seeing a sequence of vocabulary items (the more surprised, the worse the model fits the data).

We specifically focus on character-based language models for two reasons. First and foremost is the simplicity of character-based tokenization, which involves none of the assumptions about sequence groupings required by other tokenization methods. Secondly, there are a number of morphemes in Nahuatl that are written with a single character, such as the past-tense prefix *o:-*[4], some realizations of the third-person singular object prefix *k-*, and the singular-subject future tense suffix *-s*. Since these single-character morphemes are linguistically important, and subword tokenization methods risk merging them with arbitrary adjacent characters, character tokenization is more appropriate.

2.3 Related Work

There has been a great deal of research into computational approaches for assessing similarity/intelligibility between related languages and language variants, most notably highlighted in the *Workshop on NLP for Similar Languages, Varieties and Dialects* (VarDial) (Gaman et al., 2020). Particularly relevant to the work presented in this paper, Gamallo et al. (2017) describe a method for discriminating between similar languages using word and character n-gram language model perplexity. Character language models have also been shown to be effective in distinguishing between dialects of written Arabic (Sadat et al., 2014; Malmasi et al., 2015).

With respect to Nahuatl, Farfan (2019) analyzed contemporary written Nahuatl variants for points of convergence using a finite-state morphological analyzer built from a grammar of Classical Nahuatl. Other efforts in developing language technology for Nahuatl include a large parallel Nahuatl-Spanish text corpus (Gutierrez-Vasques et al., 2016), and a morphological analyzer for the Western Sierra (`nhi`) variant (Pugh et al., 2021).

3 Data

The most widely available corpus of text in the variants of Nahuatl is the Bible. We used translations into 10 different Nahuatl variants available from `scriptureearth.org`[5]. The complete list of variants employed in this study is: `azz` *Highland Puebla*, `ngu` *Guerrero*, `nch` *Central Huasteca*, `nhe` *Eastern Huasteca*, `nhy` *Northern Oaxaca*, `ncj` *Northern Puebla*, `nhi` *Western Sierra*, `nsu` *Sierra Negra*, `ncl` *Michoacán*, `nhw` *Western Huasteca*.

As translators merge verses differently in different languages, to maintain data parity for all of the variants being investigated we only included verses which were present in all variants (7,363 verses).

3.1 Orthography

Nahuatl is commonly written in a range of different orthographies. Phonemes /k/, /w/, and /h/ typically have variable graphemic representations in different orthographies. Vowel-length, which is phonemic in many Nahuatl variants but has a low functional load, can be written but is commonly ignored. See de la Cruz Cruz (2014) for a more in-depth discussion of Nahuatl orthography.

The different translations of the Bible do not adhere to a single orthographic norm, so we decided to normalize them to remove the choice of orthography as a confounding factor. Our normalization

[4] The *augment*, as it is often referred to in the literature, /o:/- is not morphologically a prefix, but is typically written attached to the verb. See Chapter 8.8 of Launey and Mackay (2011) for a detailed description of its morphological status and behavior.

[5] In fact, scriptureearth.org has translations in 11 variants, but due to an error during processing, we excluded Isthmus-Mecayapan Nahuatl (`nhx`). We plan to evaluate `nhx` in future work

Word	Segmentation	Gloss	Language Code
quinilij	∅-quin-ilij	s3sg-i3pl-tell	`azz`
oquiniluic	o-∅-quin-iluic	pst-s3sg-i3pl-tell	`ncj`
okinmilvi	o-∅-kinm-ilvi	pst-s3sg-i3pl-tell	`nsu`
oquimiluih	o-∅-quim-iluih	pst-s3sg-i3pl-tell	`nhi`

Table 1: The different forms of the ditransitive verb “to tell/say (s.t. to s.o.)” from 4 of the variants studied. Note the variation in the use of a past-marking *o-* prefix, verb stem and object prefix, and different orthographies. These forms came from Matthew 14:2, and correspond to the phrase ‘said unto (his servants)’ in the King James Bible: “And said unto his servants, This is John the Baptist;”.

method makes the following changes to account for well-known orthographic variation in contemporary Nahuatl writing:

- Replaces *hu*, *uh*, and *w* with *u*;
- Replaces *qu* followed by front vowel and *c* followed by back vowel or consonant (except *h*) with *k*;
- Neutralizes vowel length.

3.2 Language codes

For two of our three case studies, we compare our system’s analysis of Nahuatl variants with fieldwork. Since each Bible translation is associated with an ISO-639 code, and in many cases the mapping of towns/locales described in fieldwork to the variants indicated by ISO-639 codes is not clear-cut, we needed to match the ISO codes in our corpus to the variants described in the literature. To do this, we (1) consulted Ethnologue (Eberhard et al., 2021) for towns and municipalities associated with each ISO code, (2) searched for matching locations in the respective fieldwork descriptions, and (3) consulted a map to identify the closest matching place name in cases where there were no exact location matches. For more details, see Appendix A.

4 Methods

In order to analyze the three case studies described below, we evaluated the cross-variant perplexity of character language models for each Nahuatl variant in our corpus. Specifically, we split the data by verse into train (6,258 verses), dev (552 verses), and test (552 verses) partitions. For each variant, we trained a character language model on the training data for 50 epochs (this was manually verified to be sufficient for convergence). The epoch with the lowest perplexity on the dev set was selected, and the perplexity of the model at that epoch on the test set was calculated for all variants. We used PyTorch (Paszke et al., 2019) to train a unidirectional LSTM language model with 100-dimension character embeddings (with dropout) and a single recurrent layer with 1024 hidden units.

5 Case studies

In this section we present three case studies using character-based language models.

5.1 Variant identification

In order to test the usefulness of character language models for predicting the variant of a text, we combined the test set verses for all variants and calculated the perplexity for each variant’s language model on the entire data set. To produce predictions, for each verse we simply chose the variant with the lowest perplexity.

This approach yields near-perfect results (accuracy=0.99). The few errors were confusions between the different Huasteca variants, (`nhw`, `nhe`, and `nch`). This is unsurprising given their high similarity. In fact many of the verses our system incorrectly identified were identical to the same verse in the correct variant. The near-perfect performance is likely due to the restricted domain of our corpus, and the fact that the same translator(s) produced all of the verses for a given variant. Thus, it is likely that many of the patterns exploited by the language models are not language-specific (e.g. presence or absence of the *o-* prefix in the preterite) but author/document/domain specific (e.g. stylistic decisions such as word choice).

5.2 Sub-classification and genetic similarity

There are a number of different systems for subclassification of Nahuan languages. Lastra (1986), in an analysis based on synchronic lexical and grammatical similarities in 93 surveyed locations, suggests grouping Nahuatl variants into four groups: “Central”, “Huasteca”, “Western Periphery” and “Eastern Periphery”.

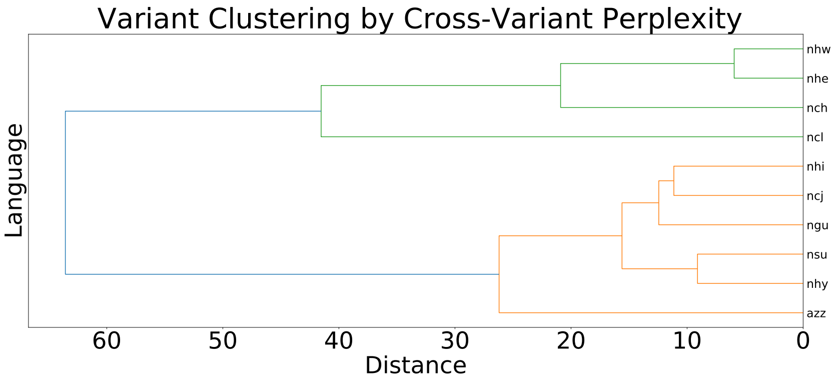

Figure 1: A dendrogram showing the variants studied, hierarchically clustered by relative perplexity. Our character language-modeling approach appears to be quite well-suited for capturing synchronic linguistic similarities between Nahuatl variants, but is less effective at identifying historical, genetic variant relationships.

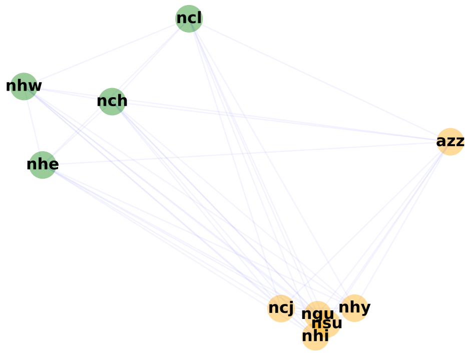

Figure 2: A force atlas diagram showing relative perplexity. Longer edges indicate higher perplexity. Node color corresponds to the clusters in Figure 1

The "East-West Split" (Canger and Dakin, 1985; Canger, 1988; Pharao Hansen, 2014) is a widely-held grouping of Nahuatl variants based on historical evidence of two waves of migration of early Nahuatl-speakers to Mexico. The first wave is thought to have resulted in what are known as the "Eastern" variants (the Huasteca and Highland Puebla variants among others), and the second in the "Western" variant group (including variants spoken near present-day Mexico City, Northern Sierra Puebla, Southeastern Puebla, and Michoacán). Importantly, many of the measurable indicators of similarity in the above two groupings, such as the existence of lexical cognates and phonological/morphological isoglosses, are often recoverable from the written form.

We grouped the variants by hierarchically clustering the vectors of cross-variant perplexity.

Central-Periphery Clustering based on the cross-variant perplexity shows a general correspondence to the Central-Periphery grouping of Lastra (1986), with some exceptions. Lastra's Central group is prominently represented in both Figure 1 (the orange lines, with the exception of the outermost `azz`) and Figure 2 (the cluster of nodes at the bottom right). The Huasteca group also stands out in our data as a cluster of three variants (`nhw`, `nhe`, `nch`) distinct from the Central group. In fact, of all variant-pairs in our data, Eastern Huasteca (`nhe`) and Western Huasteca (`nhw`) have the lowest cross-variant perplexity (clearly illustrated in Figure 1). The two Periphery groups, Western Periphery and Eastern Periphery, were not represented by any clear grouping in the cross-variant clustering, other than being separate from the *Central* group. This may be due to the lack of representation of these groups in our dataset, with only one variant from the Eastern Periphery (`azz`), and one from the Western Periphery (`ncl`).

East-West Split The distinction between Eastern and Western variants is less pronounced when clustering on cross-variant perplexity, though the distinction between the Huasteca variants and Central variants mentioned above does overlap substantially with the East-West split. The variants whose position in our grouping most contradicts the East-West sub-classification are `ngu`, `azz`, and `ncl`[6]. One possible explanation for a lack of clear distinction between the East and West groups is the fact that certain variants may tend to be more "innovative" than others, leading to new linguistic forms that set them

[6] As Pharao Hansen (2014) points out, the status of Guerrero variants within the "East-West" grouping remains unclear.

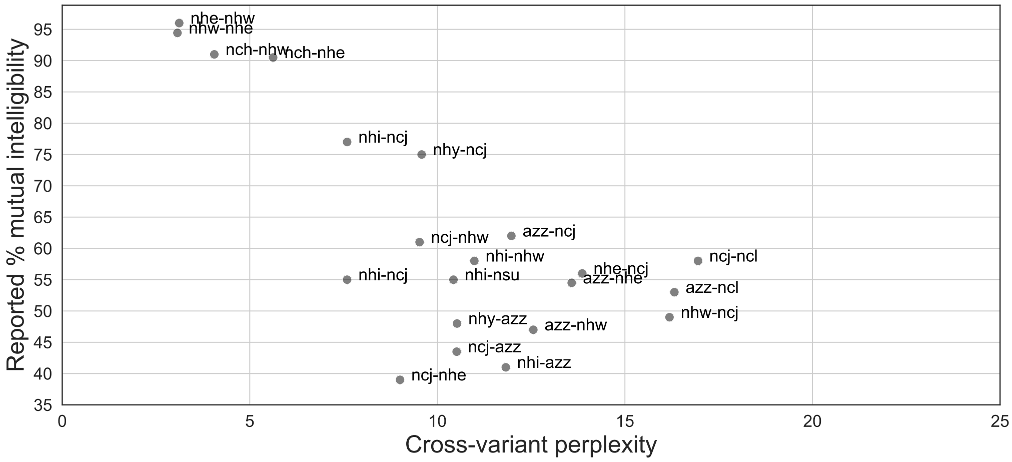

Figure 3: A plot of mutual intelligibility of variant-pairs and the corresponding cross-variant perplexity.

apart from otherwise related variants.

5.3 Mutual intelligibility

The primary systematic study of mutual intelligibility between Nahuatl varieties is Egland and Bartholomew (1978), which involved surveying speakers from 58 different communities throughout Mexico. Mutual intelligibility was assessed by playing a recording of a narrative by a speaker from a different community and asking the listener a series of comprehension questions. The results were adjusted and reported as percentages[7].

The resulting mutual intelligibility numbers are reported for community-pairs (e.g. “Tetlalpan-Xochiatipan: 99%”). In order to compare these numbers to our variant models, we assigned each community to an ISO-639 code as described in 3.2, giving us code pairs (“`nhw-nhe`: 99%”).

To evaluate whether our character language models can tell us something about mutual intelligibility, we compared each available ISO code pair’s mutual intelligibility percentages with the corresponding cross-variant perplexity. We essentially treat our language model as if it were a speaker, such that (in keeping with the above example) to compare the understanding of an `nhw` speaker listening to a narration from an `nhe` speaker, we take the language model trained on `nhw` Bible translation and evaluate its perplexity on `nhe` Bible translation. When a single language code contained multiple measurements, we used the average.

The results of this comparison, which includes all relevant measurements from Egland and Bartholomew (1978) as well as any additional reported intelligibility numbers from Ethnologue[8], are plotted in Figure 3. We found the reported mutual intelligibility between two variants and their cross-variant perplexity to be moderately negatively correlated in our data, r(19) = -0.734, p = .0002. The relationship is particularly strong for the variants with the lowest cross-variant perplexity (the Huasteca variants). However, this method is less effective at distinguishing between the mutual intelligibility of less similar variants as seen by the bunching in the center of the graph.

6 Concluding remarks

Our three case studies suggest that a simple character language model can capture a non-trivial amount of information about some of the linguistic properties, relationships, and similarities of written Nahuatl variants. The experiments also support existing literature on the utility of character features in the computational modeling of similar languages. We note the limitations of our data set, i.e. that each variant is represented by a parallel text published by the same organization (and likely by a single author per variant), and that our approach may not yield similar results on non-parallel or comparable text.

We are also interested in exploring language models under various tokenization schemes, such as unsupervised subword tokenization and morphological segmentation.

[7] We recommend consulting the first two sections of this work for details about arriving at the final percentages.

[8] Measurements reported with less than 5 speakers were excluded. Two of the measurements, `nhi-nsu` and `nhi-ncj`, were reported as “50-60%” in Ethnologue. For these data points we used 55%.

References

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, pages 893–899.

Una Canger. 1988. Subgrupos de los dialectos nahuas. In J. Kathryn Josserand and Karen Dakin, editors, *Smoke and Mist: Mesoamerican Studies in Memory of Thelma D. Sullivan*, volume 402 of *BAR lnternational*, pages 473–498. BAR, Oxford.

Una Canger and Karen Dakin. 1985. An inconspicuous basic split in nahuatl. *International Journal of American Linguistics*, 51(4):358–361.

Huadong Chen, Shujian Huang, David Chiang, Xinyu Dai, and Jiajun Chen. 2018. Combining character and word information in neural machine translation using a multi-level attention. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1284–1293, New Orleans, Louisiana. Association for Computational Linguistics.

Victoriano de la Cruz Cruz. 2014. La escritura náhuatl y los procesos de su revitalización. *Contribution in New World Archaeology*, 7:187–197.

Ted Dunning. 1994. Statistical identification of language. Technical Report 94-273, New Mexico State University.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2021. *Ethnologue: Languages of the World. Twenty-fourth edition*. SIL International. Online version: `http://www.ethnologue.com`.

S. Egland and D. Bartholomew. 1978. La inteligibilidad inter-dialectal en mexico: Resultados de algunos sondeos. Technical report.

J.I.E. Farfan. 2019. *Nahuatl Contemporary Writing: Studying Convergence in the Absence of a Written Norm*. University of Sheffield.

Pablo Gamallo, José Ramom Pichel Campos, and Inaki Alegria. 2017. A perplexity-based method for similar languages discrimination. In *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects (VarDial)*, pages 109–114.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, et al. 2020. A report on the vardial evaluation campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*. International Committee on Computational Linguistics.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for spanish-nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214.

INALI. 2009. *Catálogo De Las Lenguas Indígenas Nacionales: Variantes Lingüísticas De México Con Sus Autodenominaciones Y Referencias Geoestadísticas*. Instituto Nacional de Lenguas Indígenas, México, D.F.

Artur Kulmizev, Bo Blankers, Johannes Bjerva, Malvina Nissim, Gertjan van Noord, Barbara Plank, and Martijn Wieling. 2017. The power of character n-grams in native language identification. In *Proceedings of the 12th workshop on innovative use of NLP for building educational applications*, pages 382–389.

Yolanda Lastra. 1986. *Las areas dialectales del nahuatl moderno*. Universidad Nacional Autónoma de México, Instituto de Investigaciones Antropológicas.

M. Launey and C. Mackay. 2011. *An Introduction to Classical Nahuatl*. Cambridge University Press.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Shervin Malmasi, Eshrag Refaee, and Mark Dras. 2015. Arabic dialect identification using a parallel multidialectal corpus. In *Conference of the Pacific Association for Computational Linguistics*, pages 35–53. Springer.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Magnus Pharao Hansen. 2013. Nahuatl in the plural: Dialectology and activism in Mexico. In *Proceedings of the American Anthropological Association, Annual Meeting*.

Magnus Pharao Hansen. 2014. The East-West split in Nahuan dialectology: Reviewing the evidence and consolidating the grouping. In *Friends of Uto-Aztecan Workshop*.

Robert Pugh, Francis Tyers, and Marivel Huerta Mendez. 2021. Towards and open source finite-state morphological analyzer for zacatlán-ahuacatlán-tepetzintla nahuatl. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, pages 80–85.

Fatiha Sadat, Farzindar Kazemi, and Atefeh Farzindar. 2014. Automatic identification of arabic language varieties and dialects in social media. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 22–27.

PV Veena, M Anand Kumar, and KP Soman. 2018. Character embedding for language identification in hindi-english code-mixed social media text. *Computación y Sistemas*, 22(1):65–74.

A Language variants

In Table 2 we give the equivalences between ISO-639 language codes, variant names and locations where the variant is reported to be spoken.

Code	**Variant**	**Locations**
`azz`	Highland Puebla	Chichiquila Tatóscac Zacatipan Zautla
`nch`	Central Huasteca	Las Balsas
`ncj`	Northern Puebla	Cuaohueyalta Masacoatlán Tlaxpanaloya Xaltepuxtla
`ncl`	Michoacán	Pómaro
`ngu`	Guerrero	Copalillo Zitlala
`nhe`	Eastern Huasteca	Cuautenáhuatl Ixcatepec Jaltocan Xochiatipan Yahualica
`nhi`	Western Sierra	Tlalitzlipa
`nhw`	Western Huasteca	Casotipan Macuilocatl Tampacan Tetlalpan
`nhy`	Northern Oaxaca	—
`nsu`	Sierra Negra	—

Table 2: A listing of the locations tested for mutual intelligibility in Egland and Bartholomew (1978) as assigned to specific variants and language codes. The variants `nhy` and `nsu` did not appear in the report, but mutual intelligibility scores are available from Ethnologue (Eberhard et al., 2021).