Wancho Dialectometry: Community-created data and the Living Dictionaries project

Kellen Parker van Dam
Universität Passau
Passau, Germany
kellenparker.vandam@uni-passau.de

Abstract

Community-created lexical resources for under-documented languages represent an underexplored data source for computational dialectology. This study evaluates the viability of such data for dialectometric analysis, using the Wancho LivingDictionaries project as a case study. Wancho is a Tibeto-Burman language of the Southwestern Patkaian branch, spoken primarily in Longding District, Arunachal Pradesh, India. The dictionary is notable for being entirely community-built and speaker-facing, and uniquely among resources for Northeast India, it incorporates dialect-specific forms spanning village-level geolects and clanlects. We extract dialectal data via automated web scraping and apply a series of preprocessing steps to address inconsistencies in transcription, language labelling, and concept assignment. Pairwise linguistic distances are then computed using Sound Class Alignment (SCA, List 2010), which captures phonological similarity more accurately than raw edit distance by incorporating articulatory feature structure. The resulting distance matrix is analysed through UPGMA hierarchical clustering and NeighborNet split network inference. Despite the dataset's uneven dialect coverage and absence of systematic cognate coding, SCA-based distances recover the traditional Upper / Lower / Middle Wancho distinction and correctly situate transitional varieties. These results hold even for dialects with as few as a dozen attested forms. We show that unlike Bayesian phylogenetic inference which is poorly suited to data of this density and distribution, SCA proves to be a reliable metric. Our findings suggest that SCA distance is robust to the kinds of noise and sparsity characteristic of community-generated lexical data, and that such resources constitute a viable, if imperfect, input for automated dialectometric workflows — particularly in contexts where fieldwork-based data collection is not currently feasible.

Keywords: dialectometry, citizen science, Wancho, Tibeto-Burman

1. Introduction

This study investigates the usability of community-created dialectal data of an under-documented language in order to determine the degree to which such datasets can be used for reliable dialectometry. For this purpose, we focus on the Wancho language (Glottocode wanc1238, Hammarström et al., 2025), a Tibeto-Burman language spoken in Northeast India. Specifically, we look at the Wancho dictionary online at LivingDictionaries.app, a platform maintained by the Living Tongues institute. We focus on Wancho for a few reasons.

First, the online Wancho dictionary was created and maintained by the community themselves, despite being now largely abandoned. While linguists and anthropologists have been involved in work on Wancho for some time, the dictionary is the work of speakers working directly on their language rather than outside linguists. The dictionary is entirely community facing; while the definitions of words may be in English or Hindi, example sentences in Wancho are also often given, and headwords are provided only in the Wancho script. The script, known locally as *wan$_1$ʧo$_1$ ŋau$_2$təm$_1$* 𞋀𞋁𞋂𞋃 𞋄𞋅𞋆𞋇, is a phonemic script developed by Banwang Losu starting in 2001 (Losu 2013; 2021). It accounts for the full range of phonemic contrasts, including those not present in the dialect of the creator (Losu and Morey, 2023). However, without being able to read the script, one would not be able to use the dictionary.

Second and more importantly, the dictionary incorporates dialectal forms for the given concepts with clear phonemic transcription.[1] While Wancho is broadly divided into Upper Wancho and Lower Wancho, with an occasional mention of Middle Wancho, there are a number of small divisions within the language, including village dialects and one attested clan lect for speaker from the Khã clan. The Wancho *LivingDictionaries* project is one of only a few such dictionaries which embraces dialectal forms rather than attempting to establish or follow a standard. It is the only such case for Northeast India known to the authors.

The third reason for investigating the language through this resource is that the internal makeup of the Wancho dialects is not entirely straightforward. Wancho is an important language for understanding the internal diversity of the Patkaian languages. It is the only Patkaian language other than Chang,

[1] The Wancho alphabet specifically takes into account phonemic distinctions including those found only in certain dialects. As a result, the script can be perfectly converted to an IPA-based phonemic transcription, as was done for this study.

Proceedings of Workshop on Dialects in NLP — A Resource Perspective (DialRes) , pages 116–123
16 May 2026.

another Southwestern variety, which preserves /k/-initial pronouns which have been lost elsewhere in the group (Jacques, 2007). Thus a better understanding of its makeup would be of benefit more broadly. It is also more readily accessible than some other closely related languages due to willingness of the community to engage in work such as pedagogical discussions around dictionary creation (van Dam, 2023).

Finally, Wancho is investigated here due the author's considerable experience with the language and the Patkaian language branch within Tibeto-Burman to which it belongs. Such experience is necessary for assessing the outcomes of computationally based analyses. Without such, certain shortcomings in the method may go unnoticed if applied to a language with which one were not already somewhat familiar. This paper presents the first detailed tree model for the Wancho language.

1.1. Language Background

Wancho is a Tibeto-Burman language belonging to the Southwestern branch of Patkaian (van Dam, 2026). Little has been written on the language. Das (1988) provides a brief wordlist and scattered lexical items throughout a sociological discussion. Brief linguistic discussions were produced, including some grammatical notes (Das Gupta, 1979; Burling and Wangsu, 1998), but little if any dialectal information is clear from either on it's own.

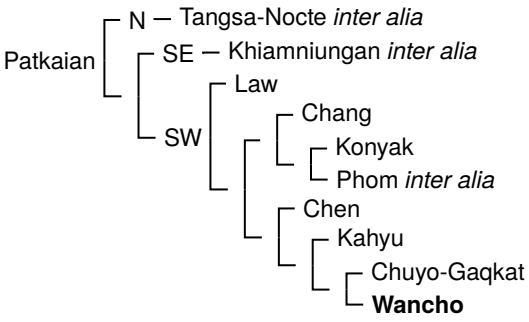

Figure 1: Place of Wancho in Patkaian

Wancho is a tonal language, with four tones based on phonation distinctions in proto Patkaian (van Dam, 2025). The typological profile is similar to closely related languages such as Konyak and Khiamniungan. There is no grammatical gender or number, although both can be explicitly marked by additional morphemes.

It is spoken almost entirely in Longding District, Arunachal Pradesh. The closest related languages are Chuyo and Gaqkat, spoken across the border in Myanmar. Some Wanchos consider these to be dialects of Wancho, and based on lexical and phonological similarities, they would be more reasonably considered as a branch of Upper Wancho. Figure 2 shows the geographic locations of the three dialect groups as traditionally given.

There is considerable dialectal diversity within Wancho, both in terms of phonology and lexicon. There are both geolects and clanlects, as well as a conventional distinction between what is called Upper Wancho and Lower Wancho. An additional Middle Wancho label is occasionally used by speakers.

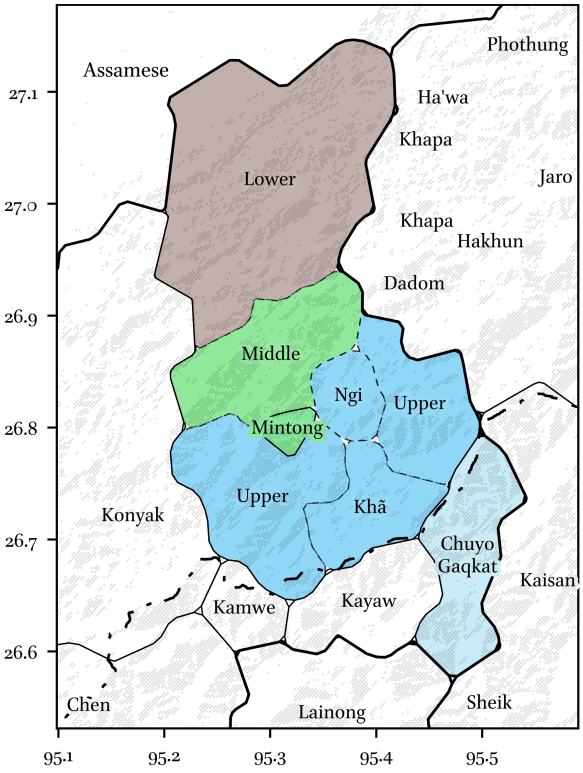

Figure 2: Locations of traditional macro groups along the Burmese border

An additional variety which falls under Wancho in terms of tribal identity is that of Mintong. It is considerably different, including its tone system and basic vocabulary such as *au* for 'mother' rather than the usual Patakaian *ɲu*, and traditional narratives hold that the residents of Mintong village settled before the arrival of the Wanchos (Losu and Morey, 2023).

Outside of the dictionary used for this study, there are two published sources of reliable lexical data the language, capturing two distinct dialects, being the wordlist of Burling & Wangsu (1998) and a follow up by Losu & Morey (2023).

2. Source data

The primary source data used for this study is scraped directly from the *LivingDictionaries* website. While the dictionary contains words from multiple dialects, many are poorly represented, with

some having only a single word. Figure 3 shows the share of each dialect, where those varieties with under 25 entries have been combined under "Others".

The majority of words in the dictionary are listed under simply "Upper Wancho" or "Lower Wancho", macro groups under which a large number of village dialects will fall. A smaller "Middle Wancho" group is also listed, although with far fewer forms given. In many cases, a single entry has been added for both a village dialect and whichever of these three macro groups under which the dialect would fall, creating multiple forms under some of the macro group labels where villages that are traditionally considered as falling under a single label differ in pronunciation. Importantly, the presence or absence of a form for a given dialect is purely of a reflection of a speaker from that dialect group adding that word.

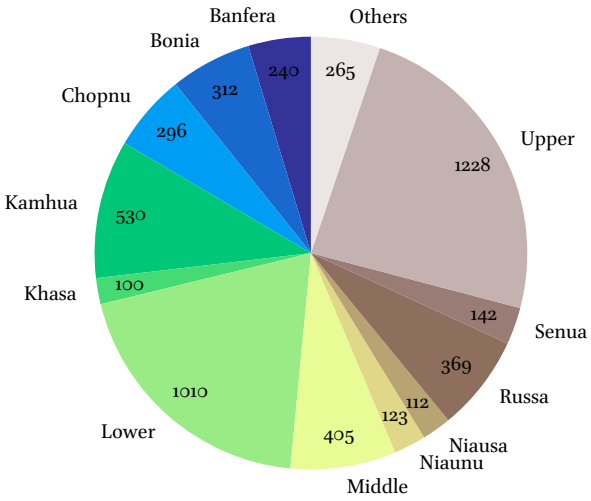

Figure 3: Dialect coverage by number of entries

We gather this data with a Python (2024) script using the *Selenium*[2] package due to the asynchronous nature of the site. The resulting data are saved as a tab-separated values (TSV) text file for further processing

2.1. Data processing

After scraping the initial dataset, some preprocessing is required prior to any analysis due to errors and inconsistencies in the data. These include example sentences in place of the English definitions, Hindi definitions in the English column, and forms which have no English entry at all but which are still recoverable based on our familiarity with the language. As an example, *ꠟꠥꠟ apa* is given under the kinship category for Middle Wancho, but with no definition. We know this to be the word for 'father', and so have corrected the entry in our processing. All such corrections were based on prior familiarity with the language, and then confirmed with a native speaker prior to any further analysis.

Another issue is around the inconsistency of language names. To resolve this, some doculect names are merged. For example, we combine "Hasse Russa" and "Russa" under one label, as these are either references to the same village, or references to two villages which are directly next to each other.[3] This provides a slightly better coverage for the varieties. We only do this in cases where it is certain that we are looking at one single geolect. There are also 252 entries for which no language variety is mentioned. We omit those from our analysis entirely.

For the remainder of the entries, typos in language names were corrected and forms were normalised. In many cases, multiple dialect names will be given in a single row, and are therefore flattened by separating these out so that reach row has only one dialect. We also flatten based on multiple forms given for a single word, although this is less common.

Headwords are then converted to IPA through the script, which can be done easily due to the phonemic nature of the orthography. We do not retain tones in the conversion, as they do not contribute to an analysis based on sound classes and are anyway cognate across varieties.[4] We then use the LingPy *ipa2tokens* function (List et al., 2018) to segment the IPA forms into space-separated tokens which can later be used for automatic detection of cognates using sound class alignment (SCA, List, 2012) as well as in tools such as Edictor (List and van Dam, 2024). This provides us with the final dataset for analysis.

A possible avenue for future research is to investigate the effectiveness of SCA distance on data which are only consisting of cognates, work which is currently being undertaken by the author for another language group. We do not take that step here, and instead only consider raw distances between forms provided for concepts as given. While results might be improved by only comparing across cognates, the purpose of this study is to assess the usefulness of the data with minimal

[2] github.com/SeleniumHQ/Selenium

[3] Three villages with Russa are given, being Hasse Russa, Kamkuh Russa, and Russa, belonging to the same dialect group. Kamkuh Russa and Russa are only a short walk apart, while Hasse Russa is just downstream from these two.

[4] While popular tools for computer-assisted language comparison such as LingPy (List et al., 2018) can handle tonal notation following the system of Chao (1930), we don't have contour data for each dialect. Either way, they would not add anything to the analysis undertaken here.

intervention. Using LingPy's LexStat cognate detection would not work in this case, as it relies on a large enough dataset to be able to determine correspondences versus coincidences. In the data analysed here, we have many concepts where only a handful of forms are given. It therefore lacks sufficient density of data. We actually can be certain of cognacy, however, such as that reflected in Table 1, due to familiarity with the language and having previously reconstructed various mesolects within proto Patkaian, to include proto Wancho (in preparation). However, this is based on knowledge and familiarity with the language and closely related languages which is not captured in the LivingDictionaries dataset.

DOCULECT	SCRIPT	FORM	TOKENS
Kamhua	𞋂𞋀𞋉𞋘𞋤	*ban maŋ*	b a n + m a ŋ
Upper	𞋂𞋀𞋉𞋘𞋤	*ban maŋ*	b a n + m a ŋ
Khasa	𞋅𞋀𞋉𞋘𞋤	*gan maŋ*	g a n + m a ŋ
Upper	𞋅𞋀𞋉𞋘𞋤	*gan maŋ*	g a n + m a ŋ
Chopnu	𞋓𞋀𞋉𞋘𞋤	*vən maŋ*	v ə n + m a ŋ
Banfera	𞋒𞋀𞋉𞋘𞋤	*wan məŋ*	w a n + m ə ŋ
Lower	𞋒𞋀𞋉𞋘𞋤	*wan məŋ*	w a n + m ə ŋ

Table 1: Sound class alignment for 'dream'

An alternative method would be to use SCA distance, as used by Edictor for cognate detection, but as can be seen in Tables 1, regular sound changes in Wancho and Patkaian more broadly frequently result in major changes of sound class. Thus, SCA alone does not produce accurate cognate judgements. We therefore forego analysing the data based on cognates, as SCA distance alone will still capture much of the signal that cognate identification would otherwise provide for such a small dataset.

2.2. Analysis

We model inter-dialect similarity using sound-class–based phonetic alignment distances computed with the SCA model (List, 2012) implemented in LingPy (List, 2012; List et al., 2018). IPA forms were tokenised into phonological segments and aligned using feature-informed scoring.[5] We used SCA distance rather than raw character edit distance (normalised Levenshtein distance) to better capture cognacy between forms. SCA distance is calculated similar to Levenshtein distance, but with articulatory features taken into consideration.

The SCA distance is calculated by reducing sounds (IPA glyphs representing segments) into broad phonetic categories called sound classes. Velar consonants form one class, as do labials, as to Alveolar consonants, and so on.

	raŋ	*zaŋ*	*gaŋ*
raŋ	0.000	0.286	0.429
zaŋ	0.286	0.000	0.429
gaŋ	0.429	0.429	0.000

The algorithm then compares these sequences by looking up the relative cost of changing one sound class into another based on a predefined scoring matrix in the form of a directed weight graph (List, 2010, p. 39). In this matrix phonetically similar classes are inexpensive to swap and dissimilar ones are expensive. Adding or removing a sound has a greater cost than simply changing a sound to a related one. This score is then normalised into a distance between 0 and 1, where 0 represents a perfect phonetic match and 1 represents no shared phonetic traits at all. We do not make any adjustments to the weights already defined in LingPy. Table 2 shows the sound classes for the stem 'sky' *#ʐaŋ* a common Patkaian etymon.[6]

In this case, *raŋ* and *zaŋ* have a shorter distance than either do to *gaŋ*, as shown in the SCA distance matrix in Table 2.2.

DOCULECT	SCRIPT	FORM	TOKENS
Khasa	𞋅𞋀𞋤	*gaŋ*	g a ŋ
Upper	𞋅𞋀𞋤	*gaŋ*	g a ŋ
Kamhua	𞋅𞋤	*gəŋ*	g ə ŋ
Upper	𞋅𞋤	*gəŋ*	g ə ŋ
Chopnu	𞋑𞋤	*zaŋ*	z a ŋ
Senua	𞋑𞋤	*zəŋ*	z ə ŋ
Lower	𞋑𞋤	*zəŋ*	z ə ŋ
Lower	𞋗𞋤	*raŋ*	r a ŋ
Russa	𞋗𞋤	*raŋ*	r a ŋ

Table 2: Sound class alignment for 'sky'

Variation between /r/ and /z/ onsets is a dialectal feature by village, but coded for Lower Wancho in both ways, as the Lower Wancho forms were coded by the same contributors who added the village-specific forms. As Hasse Russa is considered a Lower Wancho dialect speaking area, the same form was added for both at the same time. This is also an important reason why forms for Upper, Middle or Lower Wancho can not be assumed to occur as such in unlisted dialects, as the macro

[5]This is done in order to ensure we are comparing similar positions within the words, and as a typical step in the workflow for using Edictor. This step is not strictly necessary, but done in order to ensure the data are more readily available for additional analyses beyond the scope of the current study.

[6]Scriptural variations in Table 2 are due to the way in which velar codas carry the nucleus on the coda consonant. The letter 𞋤 with a dot is *-əŋ* while 𞋤 without is *-aŋ*, but then also 𞋁 is /ə/ and 𞋀 is /a/. The inclusion of the vowel is thus redundant in those forms which have it.

group forms are often just dialectal forms being over-applied.

Using the SCA analysis, pairwise dialect distances were calculated as the mean normalised alignment distance across shared lexical items. We generate a square matrix representing the linguistic distance between every pair of dialects. For every pair, it identifies the set of shared lexemes (synonyms) and calculates the average SCA distance across those forms. The resulting distance matrix was analysed using hierarchical clustering based on the unweighted pair group method with arithmetic mean (UPGMA, Sokal and Michener, 1958) and NeighborNet split network inference (Bryant and Moulton, 2004) to visualise both hierarchical structure and conflicting similarity signals. This approach situates the study within the tradition of computational dialectometry (Nerbonne and Heeringa, 1997; Heeringa, 2004; Wieling et al., 2014) while incorporating phonologically structured distance modelling.

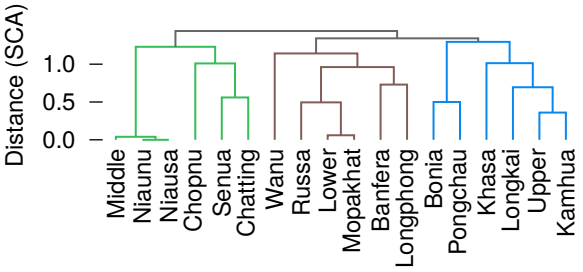

Figure 4: Neighbour joining tree with groups marked for ■ Upper, ■ Middle, and ■ Lower

The results for all dialects including the macro groups are shown in Figure 4. The tree represents the hierarchical clustering of the dialects with the inclusion of Upper, Middle and Lower. If the macro groups are removed, the overall signal of the groups is largely lost, as seen in Figure 5. This is to be expected, as in many cases, dialects such as Niau are added with Middle Wancho attached, but often as words which are not found elsewhere. Therefore, by removing the bridge between disparate concepts, we lose any sort of clear grouping which would anyway be discernible were one to simply look at the data and know the sound correspondences.

The macro groups simply have much greater representation in the source data. Additionally, different contributors have transcribed forms differently, but by adding both their version of the village's form along with the same form under one of the macro groups. Even though speakers may also add two different forms to the macro groups, such as in Table 2, the effects of such conflicting forms do not unnecessarily affect the overall connections between the macro-groups. This is discussed further below in Section 3 on data issues. We mention the results of the culled dataset here, but do not believe they are a realistic reflection of the languages' relationships given the incredibly sparse data we are left with in the absence of macro dialect labels.

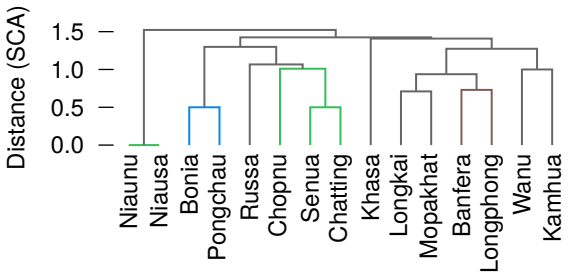

Figure 5: Neighbour joining tree without macro dialects

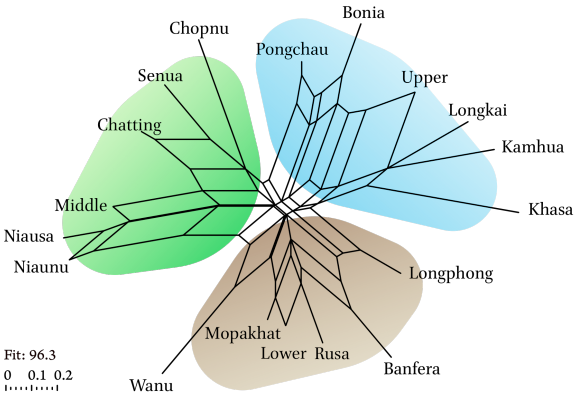

Figure 6: Neighbour net with traditional groupings of ■ Upper, ■ Middle, and ■ Lower Wancho

In general, we see the traditional three subgroups showing up well in the data, as illustrated in the neighbour joining network shown in Figure 6, in which the distances between the dialectal forms based on sound class alignments group more or less cleanly into three groups.

3. Discussion

A number of features of the data contribute to the results. First, as mentioned, many village-specific dialects are under-represented. Even after merging paired villages such as Konnu and Konsa or Niaunu and Niausa,[7] it is still not close to an even distribution.

However, this is not always a problem due to using SCA distance measures. For example, in

[7] *nu* meaning 'large' derived from 'mother' and *sa* meaning 'small' derived from 'child', a common source of augmentative and diminutive marking in the Patkaian language group to which Wancho belongs.

spite of the low number of entries for Wanu (15 items total), it is still correctly placed within Lower Wancho in the trees. This would not be the case with other approaches such as cognate coding as characters in a Bayesian phylogeny. We also see Chopnu in a stable position, as one of the more transitional varieties between Lower and Middle Wancho. Thus, we find strong signal for the Upper / Lower division frequently reported by the community, as well as a clear explanation for why Middle often comes up only occasionally as a label as a less pronounces subbranch within Lower. In cases where only one one or a few words are given, such as with Otongkhua and Lawnu, they will naturally appear as more similar to dialects for which those words are given assuming some level of shared cognacy.

For example, in Lawnu, 'give' is *koʔ*, which is also the only word for the dialect present in the dataset for that dialect. Therefore Launu will always group only with other languages which also have 'give' as *koʔ*. This will be the case regardless of where Lawnu would actually be rightfully positioned in the tree. This is especially problematic since we know that in fact *all* dialects mentioned in the LivingDictionaries dataset will have this same stem for 'give', as is the case for most of Patkaian beyond Wancho. Cognates found in related non-Patkaian groups as well, such as Jinghpaw *tʃoʔ* (Yuè, 2006). However, within the Wancho data, only a handful of dialects have been given the form for 'give', resulting in the inclusion of this one word having a much stronger influence than it ought to in terms of determining the tree model. To resolve this, we can filter the data which is being studied to only those concepts which are more widely represented in the data, such as occurring in a minimum number of dialects. A better solution is simply to remove those dialects which only have a handful of concepts, which is what we have done here. As such, only varieties with more than 5 concepts encoded in the dictionary are considered for our analysis.

On the other hand, were we to stick to only those dialects which have considerably better representation, such as those with at least 200 concepts, we would be unlikely to get out of any other views of Upper-Middle-Lower categories. That is, the results would only show that three such groups exist, with Kamhua being strictly within Upper, as nearly all Kamhua words have an identical Upper Wancho counterpart in the dataset. Such a high cutoff does not help in clarifying the internal relationships of the dialects. Instead, the obvious long-term solution is simply to incorporate more data. However, at the time of writing, there are no active contributions to the dictionary due to limits on time of some community members, and issues around access for others.[8]

Translations or other concept issues are also problematic, with many coming directly from Google searches, and at a time when Google was plagued by AI trained on Reddit (Olson and Kerr, 2024; Tong, 2024). For example, the Chopnu dialect form for 'generation' 𞋗𞋁𞋝𞋛𞋁𞋝 instead gives a summary of a 2019 computer game called Generation Zero (2019). 'Banana' 𞋍𞋞𞋬 for Senua has as a semantic domain "Units of measure", unintentionally referencing a popular meme from the online forum *Reddit*. This is one issue with a dictionary where English is the target language for definitions but may not be widely understood within the community, and for which the contributors may rely on potentially misguided Google AI search results when those were particularly precarious.

DOCULECT	SCRIPT	FORM	TOKENS		
Nginu	𞋂𞋁𞋔	*bək*	b	ə	k
Konnu	𞋂𞋁𞋔	*bək*	b	ə	k
Konsa	𞋂𞋁𞋔	*bək*	b	ə	k
Upper	𞋂𞋁𞋔	*bək*	b	ə	k
Kamhua	𞋂𞋁𞋔	*bək*	b	ə	k
Khasa	𞋅𞋁𞋔	*gək*	g	ə	k
Upper	𞋅𞋁𞋔	*gək*	g	ə	k
Kamhua	𞋓𞋁𞋔	*vək*	v	ə	k
Upper	𞋓𞋁𞋔	*vək*	v	ə	k
Pongchau	𞋒𞋁𞋔	*wək*	w	ə	k
Upper	𞋒𞋁𞋔	*wək*	w	ə	k
Russa	𞋒𞋁𞋔	*wək*	w	ə	k
Lower	𞋒𞋁𞋔	*wək*	w	ə	k

Table 3: Sound class alignment for 'pig'

Finally, there are considerable variations in how terms are transcribed, as previously mentioned above. In some cases these are clearly due to dialectal differences. In other cases it could either be the result of intra- or inter-speaker variation, an unfamiliarity with the script, or uncertainty around the phonetic realisation of a term.

As an example, the term 'pig' is given in Table 3. For Kamhua, the term is given as both 𞋂𞋁𞋔 *bək* and 𞋓𞋁𞋔 *vək*, which is a reasonable case of allophony or the transcription of a sound which may not be entirely clear to the speakers themselves. However, the same term occurs for the Upper Wancho macrogroup as both 𞋒𞋁𞋔 *wək* and 𞋅𞋁𞋔 *gək*. The former is again understandable as

[8] It should be mentioned that, at the time of writing, a new version of the dictionary is being developed for the community with less access restrictions, and in way in which such automated analyses as the one described here would be more accessible.

allophony/allography from *vək*, whereas the /g/ onset is actually a dialectal difference, found in the Khasa dialect. However, as Kamhua and Khasa are both considered Upper Wancho dialects, the "Upper Wancho" entries reflect both pronunciations as speakers from both villages added their forms as Upper Wancho forms. A more comprehensive study would need to address these issues to provide a clearer picture of the internal relationships between dialects.

Still, it is our hope that contributions resume once more. For this reason, we do not provide a full CLDF dataset at this time.

SCA will fail on clear cognates.

4. Conclusion

The analysis presented here shows that even with a fairly sparse dataset across multiple dialects, by analysing lexical data based on sound class alignments (SCA), reliable genealogies can be produced. By limiting our analysis to only those concepts which exist in at least three doculects, we prevent strictly paired entires such as Upper Wancho and Kamhua where both were added at once for the same concepts and with identical segmental forms. By only investigating concepts with at least 5 shared concepts, we avoid unnecessary noise caused by under-represented varieties.

Importantly, however, this analysis relies on a considerable amount of work to ensure that concepts are properly linked in the code. Ensuring that forms given properly match the concepts found throughout the dialects given requires some understanding of the language, both in terms of historical phonology and lexical formation strategies.

However, armed with this knowledge, we are able to produce an otherwise automated analysis that takes community-managed data as its input, analyses a community-developed phonemic script, and as a gain quantifiable evidence for internal subgrouping. This same conclusion could be gained through a comparative study with a minimal concept list were one to travel through the district and then investigating the forms provided. However, our approach shows that even in lieu of direct fieldwork or consistent methodologies around data collection, non-academic sources of data still provide an incredibly important asset for language typology.

Finally, based on the current analysis combined with discussions with native speakers and previous analyses on the language for the purposes of cross-linguistic comparisons, we propose the first detailed tree model for Wancho in Figure 7 a tree model for Wanchoic. We include Chuyo and Gaqkat, based on unpublished fieldwork data no present in the LivingDictionaries project, which show a high degree of similarity with Upper Wancho varieties. A number of lexical differences are found, moreso in Gaqkat than Chuyo, which could either indicate strata effects caused by non-Wancho Patkaian languages spoken in Myanmar, or, that Gaqkat represents an earlier form of Southwestern which has largely shifted more toward Wancho as a result of close contact with Chuyo, which is more clearly aligned with Upper Wancho.

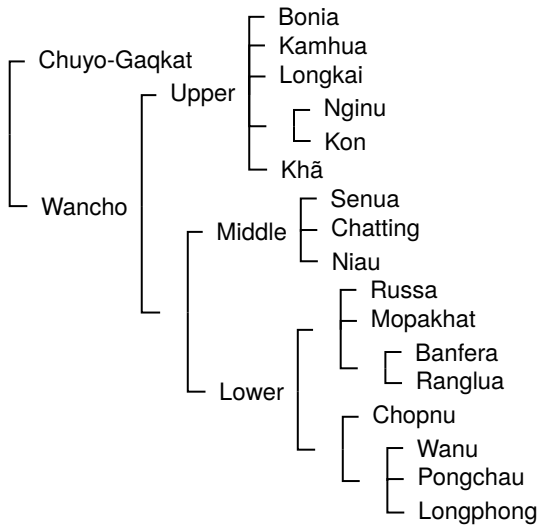

Figure 7: Internal structure of Wanchoic

This analysis is based on both the findings of the current study, along with community descriptions of language variety relatedness. Our results here show that conventional descriptions by speakers are largely accurate for Wancho, supported by both regular sound changes indicative of subbranches, as well as lexical variation for which Lower Wancho shows more similarity to neighbouring Northern Patkaian languages such as Nocte.

5. Acknowledgements

Thank you to Banwang Losu for his help confirming corrections made to the scraped data, and to the members of the Wancho Literary Mission for their work on the language. Thank you to Mattis List and Jessica Nieder for their comments on an earlier version of this paper. Finally, thank you to the anonymous reviewers for their comments and suggestions.

6. Supplemental Materials

All code and data used in this study are available on Codeberg repository at patkaiist/wancho. The original LivingDictionaries project is available at livingdictionaries.app/wancho.

7. Bibliographical References

David Bryant and Vincent Moulton. 2004. Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. *Molecular Biology and Evolution*, 21(2):255–265.

Robbins Burling and Mankai Wangsu. 1998. Wancho phonology and word list. *Linguistics of the Tibeto-Burman Area*, 21(2):43–71.

Yuen Ren Chao. 1930. ə Sistim əv Toun-letəz. *Le Maître Phonétique*, 30(30):24–27.

Pranjana Das. 1988. *Wancho through history*. Ph.D. thesis, North-Eastern University, Shillong.

Kamalesh Das Gupta. 1979. A note on the Wancho language of Arunachal Pradesh. *Resarun: journal of the Research Department, Government of Arunachal Pradesh*, 5(1):25–37.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2025. Glottolog 5.2.

Wilbert Heeringa. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. Ph.D. thesis, University of Groningen.

Guillaume Jacques. 2007. A shared suppletive pattern in the pronominal systems of Chang Naga and Southern Qiang. *Cahiers de linguistique Asie orientale*, 36(1):61–78.

Johann-Mattis List. 2010. SCA: Phonetic alignment based on sound classes. In *European Summer School in Logic, Language and Information*, pages 32–51. Springer.

Johann-Mattis List. 2012. LexStat: Automatic Detection of Cognates in Multilingual Wordlists. *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, pages 117–125.

Johann-Mattis List, Simon J. Greenhill, and Russell D. Gray. 2018. LingPy: A Python Library for Historical Linguistics. *Journal of Open Source Software*, 3(28):1242.

Johann-Mattis List and Kellen Parker van Dam. 2024. Computer-Assisted Language Comparison with EDICTOR 3. In *Proceedings of the 5th Workshop on Computational Approaches to Historical Language Change*, pages 1–11.

Banwang Losu. 2013. *The Wancho Script*. Partridge India, New Delhi.

Banwang Losu. 2021. Phonology of the Wancho Language and Script. Master's thesis, Department of Linguistics, Deccan College, Pune. MA dissertation.

Banwang Losu and Stephen Morey. 2023. The Wancho language of Kamhua Noknu village. *Linguistics of the Tibeto-Burman Area*, 46(2):201–234.

John Nerbonne and Wilbert Heeringa. 1997. Measuring Dialect Distance Phonetically. In *Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology*, pages 11–18.

Paayal Olson and Dara Kerr. 2024. Google strikes $60 million deal with Reddit, allowing search giant to train AI models on human posts. *CBS News*.

Python Software Foundation. 2024. Python 3.13. Computer software.

Robert R. Sokal and Charles D. Michener. 1958. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438.

Systemic Reaction / Avalanche Studios. 2019. Generation Zero. generationzero.com. Computer software.

Anna Tong. 2024. Why Google's AI Overviews Gets Things Wrong. *MIT Technology Review*.

Kellen Parker van Dam. 2023. Developing & Maintaining Community-Driven Language Resources: Dictionaries & Narratives. Talk presented at the Workshop on Pedagogic Development. Organised by the Wancho Literary Mission with support from the Assam Rifles, Longding, Arunachal Pradesh.

Kellen Parker van Dam. 2025. On the independence of tonogenesis in Patkaian branches. *Linguistics of the Tibeto-Burman Area*, 48(2):163–192.

Kellen Parker van Dam. 2026. Patkaian (Northern Naga). In Kristine Hildebrandt, Yankee Modi, David Peterson, and Hiroyuki Suzuki, editors, *The Oxford Guide to the Tibeto-Burman Languages*. Oxford University Press, Oxford. Forthcoming.

Martijn Wieling, John Nerbonne, and R. Harald Baayen. 2014. Quantitative Social Dialectology: Explaining Linguistic Variation Geographically and Socially. *PLOS ONE*, 9(9):e105342.

Yuè Má Là 岳麻腊. 2006. 景颇语杜连话概况 [Overview of the Duliang Dialect of the Jingpo Language]. 民族语文, (4):68–80.